

Содержание

Введение	11
1 Все, что вы жаждали знать об электронных таблицах, но боялись спросить	21
Немного данных для примера.....	22
Быстрый просмотр с помощью кнопок управления	23
Быстрое копирование формул и данных	24
Форматирование ячеек.....	25
Специальная вставка.....	27
Вставка диаграмм	28
Расположение меню поиска и замены.....	29
Формулы поиска и вывода величины.....	30
Использование VLOOKUP/ВПР для объединения данных	32
Фильтрация и сортировка	33
Использование сводных таблиц	37
Использование формул массива.....	40
Решение задач с помощью «Поиска решения».....	41
OpenSolver: хотелось бы обойтись без него, но это невозможно.....	46
Подытожим	47
2 Кластерный анализ, часть I: использование метода k-средних для сегментирования вашей клиентской базы	49
Девочки танцуют с девочками, парни чешут в затылке.....	51
Реальная жизнь: кластеризация методом k-средних в электронном маркетинге	56
Оптовая Винная Империя Джоуи Бэг О'Донатса	56
Исходный набор данных	57
Определяем предмет измерений	58

Начнем с четырех кластеров.....	61
Евклидово расстояние: измерение расстояний напрямик.....	62
Расстояния и принадлежность к кластеру для всех!.....	65
Поиск решений для кластерных центров.....	67
Смысл полученных результатов.....	70
Рейтинг сделок кластерным методом.....	71
Силуэт: хороший способ позволить разным значениям к посостязаться.....	75
Как насчет пяти кластеров?.....	82
Поиск решения для пяти кластеров.....	83
Рейтинг сделок для всех пяти кластеров.....	84
Вычисление силуэта кластеризации по пяти средним.....	87
К-медианная кластеризация и асимметрическое измерение расстояний.....	89
Использование k-медианной кластеризации.....	89
Переходим к соответствующему измерению расстояний.....	90
А теперь все то же самое, но в Excel.....	92
Рейтинг сделок для 5-медианных кластеров.....	94
Подытожим.....	98

3 Наивный байесовский классификатор

и неопишуемая легкость бытия идиотом.....	101
Называя продукт Mandrill, ждите помех вместе с сигналами.....	101
Самое быстрое в мире введение в теорию вероятности.....	104
Суммируем условную вероятность.....	104
Совместная вероятность, цепное правило и независимость.....	105
Что же с зависимыми событиями?.....	106
Правило Байеса.....	107
Использование правила Байеса для создания моделирования.....	108
Высококласные вероятности часто считаются равными.....	110
Еще немного деталей классификатора.....	111
Да начнется Excel-вечеринка!.....	113
Убираем лишнюю пунктуацию.....	113
Разное о пробелах.....	114
Подсчет жетонов и вычисление вероятностей.....	118
У нас есть модель! Воспользуемся ею.....	121
Подытожим.....	127

4 Оптимизационное моделирование:

этот «свежевыжатый апельсиновый сок» не смешает себя сам.....	129
Зачем ученым, работающим с данными, нужна оптимизация?.....	130
Начнем с простого компромисса.....	131
Представим проблему в виде политопа.....	132

Решение путем сдвигания линии уровня функции	134
Симплекс-метод: все по углам	135
Работа в Excel	137
Монстр в конце главы	147
Свежий, из сада — прямо в стакан...	
с небольшой остановкой на модель смешивания.....	148
Вы используете модель для смешивания	149
Начнем с характеристик.....	150
Возвращаемся к консистенции	151
Вводим данные в Excel	152
Постановка задачи «Поиску решения»	155
Снижаем стандарты	158
Удаление дохлых белок: правило минимакс	161
«Если... то» и ограничение «Большого M»	164
Еще больше переменных: добыем до 11	167
Моделируем риски	175
Нормальное распределение данных	176
Подытожим	184
5 Кластерный анализ, часть II: сетевые графы и определение сообществ....	187
Что такое сетевой граф?.....	188
Визуализируем простой граф	189
Краткое введение в Gephi.....	192
Установка Gephi и подготовка файлов	192
Визуализация графа	194
Степень вершины	197
Приятная картинка	200
Прикосновение к данным графа	200
Строим граф из данных об оптовой торговле вином	202
Создание матрицы близости косинусов.....	204
Построение графа N-соседства	207
Числовое значение ребра: очки и штрафные в модулярности графа.....	212
Кто же такие «очки» и «штрафные»?	212
Подготовка к итоговому подсчету.....	216
Переходим к кластеризации!	219
Деление 1	219
Деление 2: электролатино!	225
И... деление 3: возмездие	227
Кодируем и анализируем группы	228
Туда и обратно: история Gephi	233
Подытожим	238

6	Бабушка контролируемого искусственного интеллекта – регрессия	241
	Погоди, ты что – беременна?	241
	Не обольщайтесь!	242
	Определение беременных покупателей РитейлМарта с помощью линейной регрессии	243
	Набор отличительных признаков	244
	Сборка обучающих данных	245
	Создание фиктивных переменных	247
	Мы сделаем свою собственную линейную регрессию!	250
	Статистика линейной регрессии: R-квадрат, критерии Фишера и Стьюдента	259
	Делаем прогнозы на основании новых данных и измеряем результат	270
	Предсказание беременных покупателей РитейлМарта с помощью логистической регрессии	281
	Первое, что нам нужно – это функция связи	281
	Присоединение логистической функции и реоптимизация	282
	Создание настоящей логистической регрессии	286
	Выбор модели: сравнение работы линейной и логистической регрессий	287
	Дополнительная информация	291
	Подытожим	292
7	Комплексные модели: огромная куча ужасной пиццы	293
	Используем данные из главы 6	294
	Бэггинг: перемешать, обучить, повторить	296
	Одноуровневое дерево решений – неудачное название «неумного» определителя	296
	А мне не кажется, что это глупо!	297
	Нужно еще сильнее!	300
	Обучим же ее!	300
	Оценка бэггинговой модели	310
	Бустинг: если сразу не получилось, бустингуйте и пробуйте снова	315
	Обучаем модель: каждому признаку – шанс	315
	Оценка модели бустинга	324
	Подытожим	327
8	Прогнозирование: дышите ровно, выиграть невозможно	329
	Торговля мечами начата	330
	Знакомство с временной последовательностью данных	331
	Медленный старт с простым экспоненциальным сглаживанием	333
	Настраиваем прогноз простого экспоненциального сглаживания	335

Возможно, у вас есть тренд	341
Экспоненциальное сглаживание Холта с корректировкой тренда	344
Настройка холтовского сглаживания с коррекцией тренда в электронной таблице ...	346
Мультипликативное экспоненциальное сглаживание Холта–Винтерса	360
Установка исходных значений уровня, тренда и сезонности	362
Приступим к прогнозу	367
И наконец... оптимизация!	372
Пожалуйста, скажите, что это все!!	373
Создаем интервал прогнозирования вокруг прогноза	374
И диаграмма с областями для пущего эффекта	378
Подытожим	381
9 Определение выбросов: выделяющиеся не значит важные	383
Выбросы тоже (плохие?) люди!	384
Захватывающее дело Хадлум против Хадлум	384
Границы Тьюки	386
Применение границ Тьюки в таблице	386
Ограничения этого нехитрого метода	388
Ни в чем не ужасен, плох во всем	390
Подготовка данных к отображению на графе	391
Создаем граф	394
Вычисляем к ближайших соседей	397
Определение выбросов на графе, метод 1: полустепень захода	398
Определение выбросов на графе, метод 2: нюансы k-расстояния	401
Определение выбросов на графе, метод 3: факторы локальных выбросов – это то, что надо	403
Подытожим	409
10 Переходим от таблиц к программированию	411
Налаживаем контакт с R	412
Пошевелим пальцами	413
Чтение данных в R	421
Настоящая научная работа с данными	423
Сферическое k-среднее винных данных в нескольких линиях	423
Построение моделей ИИ для данных о беременных	430
Прогнозирование в R	439
Определение выбросов	443
Подытожим	448
Заключение	451
Благодарности	459

Моей жене Лидии.

*То, что ты делаешь каждый день – круто! Если бы не ты,
я бы лишился волос (и ума) миллиард лет назад*

Введение

Что я здесь делаю?

Наверняка где-нибудь в прессе, финансовой литературе и журналах или на конференции вы слышали что-то об обработке данных, их представлении и анализе — том, что составляет «науку о данных». Эта наука может предсказать результаты выборов, рассказать о ваших покупательских привычках больше, чем вы осмелились бы поведать маме, и определить, на сколько лет сокращают вашу жизнь сырные буррито с чили.

В последнее время вокруг науки о данных наблюдается некоторый ажиотаж, который начинает оказывать давление на многие виды бизнеса. Не занимаясь анализом данных, вы рискуете потерпеть неудачу в конкурентной борьбе. Обязательно появится кто-нибудь, разработавший очередной новый продукт под названием «Что-то-про-графы-и-большие-данные», — и уничтожит ваш бизнес.

Сделайте глубокий вдох.

Не все так мрачно! Вас, несомненно, спасет то, что большинство тех, кто считает себя «доками» в науке о данных, делают все ровно наоборот. Они начинают с покупки программ и нанимают консультантов. Они тратят все свои деньги еще до того, как поймут, чего же они на самом деле хотят. Заказав программные инструменты, они считают, что сделали главное и можно расслабиться.

Прочитав эту книгу, вы будете на голову выше этих «специалистов». Вы будете иметь точное представление о том, что такое техники анализа данных и как они используются. И когда придет время планировать, нанимать и покупать, вы уже будете знать, как применить возможности науки о данных с пользой именно для вашей конкретной компании.

Цель этой книги — введение в практическую науку о данных в комфортном режиме беседы. Надеюсь, что по окончании чтения священный ужас перед этим таинственным «зверем» — данными — сменится энтузиазмом и мыслями о том, как с их помощью поднять свой бизнес на новый уровень.

Рабочее определение науки о данных

В некоторой степени наука о данных — синоним таких терминов, как бизнес-аналитика; исследование операций; бизнес-интеллект; промышленный шпионаж; анализ, моделирование и раскрытие данных (также называемое обнаружением знаний в базах данных, или ОЗБД). Иными словами, нынешняя наука о данных — просто новый виток того, чем люди занимаются уже довольно долго.

После расцвета вышеозначенных и других дисциплин произошел скачок в технологиях. Совершенствование аппаратной и программной платформ сделали легким и недорогим сбор и анализ больших объемов данных во всех областях — будь то продажи и маркетинг, запросы HTTP с вашего сайта или информация для поддержки клиентов. Малый бизнес и некоммерческие организации могут теперь привлекать аналитиков, содержание которых раньше могли себе позволить только большие корпорации.

Конечно, из-за того, что наука о данных используется как всеобъемлющее ученое словечко для обозначения аналитики сегодня, она чаще всего ассоциируется с техниками добычи данных (data mining), такими как искусственный интеллект, кластерный анализ и определение выбросов. Благодаря подешевевшей аппаратной поддержке, обеспечившей резкий рост количества переменных бизнес-данных, эти вычислительные техники стали опорой бизнеса в последние годы, хотя раньше они были слишком громоздкими для использования на производстве.

В этой книге я собираюсь дать широкий обзор всех разделов науки о данных. Вот определение, которое я буду использовать:

Наука о данных — это трансформация данных методами математики и статистики в рабочие аналитические выводы, решения и продукты.

Я определяю это понятие с точки зрения бизнеса. В нем упоминается применимый и полноценный готовый продукт, получаемый из данных. Почему? Потому что я занимаюсь этим не в исследовательских целях и не из любви к искусству. Я изучаю данные для того, чтобы помочь моей компании работать лучше и постоянно повышать свою эффективность; поскольку вы держите в руках мою книгу, подозреваю, что наши намерения схожи.

Используя это определение, я собираюсь описать вам основные техники анализа данных, такие как оптимизация, прогнозирование и моделирование, а также затронуть наиболее актуальные темы — искусственный интеллект, сетевые графы, кластерный анализ и определение выбросов.

Одни из этих техник довоенные в буквальном смысле слова. Другие внедрены в течение последних 5 лет. Но вы увидите, что возраст не имеет никакого отношения к сложности или полезности. Все эти техники — независимо от степени популярности — одинаково полезны для бизнеса при правильном выборе.

Вот почему вам нужно понимать, какая техника для решения какой проблемы подходит, как эти техники работают и как их моделировать. Довольно много людей имеют представление о сути одной или двух описанных мною техник — этим их знания и ограничиваются. Если бы у меня в ящике для инструментов был только молоток, наверное, я бы пытался решать все проблемы ударом по сильнее. Совсем как мой двухлетний сын.

Но поскольку мне не два года, я предпочитаю иметь еще какие-то инструменты в своем распоряжении.

Но подождите, а как же большие данные?

Наверняка вы слышали термин «большие данные» даже чаще, чем «наука о данных». О них ли эта книга?

Ответ зависит от того, что понимать под большими данными. Если вы определяете большие данные как подсчет сводной статистики неструктурированного мусора, хранящегося в горизонтально масштабируемом NoSQL-массиве, то нет, это книга не о больших данных.

Если вы определяете большие данные как превращение переменных данных в решения и аналитические выводы с помощью ультрасовременных методов анализа (независимо от того, где хранится информация), тогда да, моя книга о больших данных.

В этой книге не рассматриваются системы управления базами данных, такие как MongoDB и Hbase. В ней не рассказывается о пакетах для разработчиков, таких как Mahout, NumPy, различных R-библиотеках и т.д. Для этого существуют другие книги.

Я сделал так намеренно. Эта книга игнорирует инструменты, хранилища и код. Вместо этого она, по возможности, фокусируется на методах. Многие думают, что если смешать хранение и извлечение данных с щепоткой очистки и агрегации, получится коктейль «Все, что нужно знать о больших данных».

Они ошибаются. Эта книга поможет вам беспрепятственно пробиться сквозь завесу многозначительной болтовни, которой нас окружают продавцы программного обеспечения для работы с большими данными и блогеры, и покажет вам, на что на самом деле способны ваши данные. Что примечательно, для большинства этих техник объем ваших данных может быть любым — крошечным или

огромным. Вы не обязаны иметь петабайт данных и энную сумму с пятью нулями на предсказание интересов вашей огромной клиентской базы. Иметь массив данных — это, конечно, замечательно, однако есть бизнесы, прекрасно обходящиеся и без этого «сокровища», более того — никому не хочется их генерировать. Например, мяснику, торгующему в моем родном квартале. Но это не значит, что его бизнесу помешало бы небольшое кластерное разделение «бекон/колбаса».

Если сравнивать книги с видами спорта, моя книга сравнима с гимнастикой. Никаких тренажеров и упражнений на выносливость. Поняв, как реализовывать техники с помощью базовых инструментов, вы обнаружите, что свободно можете применять их во многих технологиях, с легкостью моделировать их, правильно выбирать программные продукты у консультантов, формулировать задачи программистам и т. д.

Кто я?

Давайте прервемся ненадолго, и я расскажу вам о себе. Научный подход к изучению данных, который я проповедую, возник не вчера — к нему меня вел долгий путь. Много лет назад я был консультантом по менеджменту. Я работал над аналитическими проблемами таких организаций, как ФБР, министерство обороны США, компания Coca-Cola, группы отелей Intercontinental и Royal Carriibbean. Из всего этого опыта я вынес одно: наука о данных должна стать прерогативой не только ученых.

Я работал с менеджерами, которые покупали симуляции, когда им были нужны модели оптимизации. Я работал с аналитиками, которые понимали только графики Ганта*, так что абсолютно все приходилось представлять в виде этих графиков. Как консультанту, мне было нетрудно расположить к себе покупателя, имея в арсенале любые старые бумаги и миленькую презентацию в PowerPoint, потому что они не могли отличить искусственный интеллект от бизнес-анализа, а бизнес-анализ — от BS.

Цель этой книги — расширение аудитории, способной понять и применить техники научного анализа данных. Я не пытаюсь обратить вас, уважаемые читатели, в специалистов по научной обработке данных против вашей воли. Я просто хочу, чтобы вы научились применять науку о данных настолько, насколько сможете, в той области, в которой вы уже хорошо разбираетесь.

Это заставляет задать вопрос: кто же вы?

* Популярный тип столбчатых диаграмм (гистограмм), который используется для иллюстрации плана, графика работ по какому-либо проекту. Является одним из методов планирования проектов. — *Прим. ред.*

Кто вы?

Не пугайтесь, я не использовал научный анализ данных, чтобы шпионить за вами. Я понятия не имею, кто вы, но заранее благодарен вам за то, что раскошались на эту книгу.

Вот несколько архетипов (или личностей — для вас, маркетологи!), которые пришли мне на ум, когда я писал эту книгу. Возможно, вы:

- заместитель начальника по маркетингу и хотите использовать свои бизнес-переменные стратегическим образом, для оценки продукта и сегмента рынка, но не понимаете подходов, рекомендуемых разработчиками приложений и переоцененными консультантами;
- аналитик, предсказывающий спрос, который знает, что история заказов фирмы содержит больше информации о клиентах, чем даже план на следующий квартал;
- руководитель розничного интернет-магазина, желающий угадать по данным о предыдущих заказах, когда клиент скорее всего «созреет» для очередной покупки;
- бизнес-аналитик, который в состоянии просчитать растущие денежные потоки и затраты на снабжение, но не знает, как перебросить мостик экономики на издержках;
- онлайн-маркетолог, который хочет чего-то большего для своей компании от бесплатных текстовых сервисов, таких как электронные письма или социальные сети. Пока же судьба разосланных сообщений незавидна — их открывают и тут же выбрасывают в корзину.

Иными словами, вы — читатель, который получает практическую пользу от дополнительной информации о научной обработке данных, но пока не нашел «свой конек» во всем многообразии техник. Цель этой книги — стряхнуть мишуру (код, инструменты и просто слухи) с науки о данных и обучить необходимым техникам на практических примерах, понятных любому, прошедшему курс линейной алгебры или вычислительной математики в институте. Если вы, конечно, их успешно сдали. Если нет — читайте медленно и не стесняйтесь пользоваться Википедией.

Никаких сожалений — только электронные таблицы

Эта книга не о программировании. Я даже готов гарантировать полное отсутствие (ну, по крайней мере, до главы 10) в ней кода. Почему?

Да потому что я не хочу тратить первые сто страниц на возню с Git, объявлением переменных среды и выступление Emacs против Vi.

Если вы пользуетесь исключительно Windows и Microsoft Office, работаете в государственной структуре и вам запрещено скачивать и устанавливать приложения из каких попало открытых источников и даже если MATLAB или ваш графический калькулятор наводили на вас леденящий ужас во времена студенчества, вам нечего бояться.

Нужно ли вам знать, как пишется код, чтобы перевести большую часть этих техник в автоматизированную, производственную форму? Непременно! Вы или кто-то из ваших коллег должен знать технологии хранения данных и уметь управляться с кодом.

Нужно ли вам знать, как пишется код, чтобы понимать, различать и моделировать эти техники? Совершенно ни к чему!

Именно поэтому я объясняю каждую методику с помощью электронных таблиц.

Ну, ладно, если по-хорошему, то я должен признаться, что все вышесказанное мною не совсем правда. Последняя глава этой книги — о переходе на язык программирования R, ориентированный на анализ данных. Она предназначена для тех из вас, кто захочет использовать эту книгу как трамплин к пониманию новых глубин аналитики.

Но электронные таблицы так устарели!

Электронные таблицы — не самый привлекательный инструмент из существующих.

Электронные таблицы стоят немного особняком. Они позволяют вам видеть данные и взаимодействовать с ними (или, по крайней мере, кликать на них). Они создают определенную свободу для маневра. Во время изучения анализа данных вам понадобится инструмент — привычный, понятный каждому, позволяющий двигаться быстро и легко в процессе. Это и есть электронные таблицы.

Давайте, наконец, скажем себе: «Я человек и обладаю чувством собственного достоинства. Я не должен делать ручную работу программного фреймворка, чтобы научиться анализировать данные».

А еще электронные таблицы отлично подходят для прототипирования! Конечно, вы не запустите с их помощью производственную модель ИИ* для вашего интернет-магазина из программы Excel, но зато сможете понять характер заказов, спрогнозировать, какие продукты в будущем вызовут интерес потребителей, и разработать прототип модели для определения целевой аудитории.

Используйте Excel или LibreOffice

Все примеры, с которыми вам придется работать, отображаются в таблицах Excel.

На сайте этой книги (www.wiley.com/go/datasmart) размещены электронные таблицы с открытым доступом для каждой главы, так что вы сможете следить за ходом повествования. Если вы по натуре склонны к риску, можете стереть оттуда все данные, кроме исходных, и сделать всю работу самостоятельно.

Эта книга совместима с Excel версий 2007, 2010, 2011 для MacOS и 2013. В первой главе достаточно подробно рассматриваются различия между версиями.

У большинства из вас есть доступ к Excel и вы наверняка уже применяете его в вашей работе для отчетности или хранения информации. Но если по какой-то причине этой программы у вас нет, вы можете ее либо купить, либо воспользоваться бесплатным аналогом от LibreOffice (www.libreoffice.org).

А КАК ЖЕ GOOGLE DRIVE?

Кто-то из вас наверняка спросит, можно ли при решении задач, которые нам предстоят, использовать Google Drive — облачный сервис, доступный с любого устройства, как почтовый ящик. Что и говорить, вариант заманчивой... К сожалению, он не будет работать.

Google Drive отлично справляется с небольшими таблицами, но того, чем собираетесь заниматься вы, он просто не выдержит. Процесс добавления строк и колонок уже раздражает, реализация поиска решения просто ужасна, а у графиков даже нет линий тренда!

Хотелось бы мне, чтобы было иначе, но увы...

LibreOffice — открытый бесплатный ресурс, имеющий практически всю функциональность Excel. Я даже думаю, что его собственный поиск решений предпочтительнее, чем у Excel. Так что если вы не раздумали читать эту книгу — вперед!

* Искусственный интеллект. — Прим. пер.

Условные обозначения

Чтобы помочь вам извлечь из текста максимальную пользу, я ввел в эту книгу несколько условных обозначений.

ВСТАВКИ

Вставки типа той, в которой вы только что прочитали про Google Drive, раскрывают «побочные» темы, упомянутые в тексте.

ВНИМАНИЕ!

Эти разделы содержат важную информацию, напрямую связанную с текстом, которую я рекомендую все время держать в уме.

ЗАМЕТКИ

Здесь вы найдете советы, подсказки, приемы и все в этом духе, что пришлось к слову в текущем обсуждении.

Частенько я буду вставлять в текст небольшие кусочки кода Excel вроде этого:

```
=CONCATENATE("THIS IS A FORMULA", "IN EXCEL!") /  
=СЦЕПИТЬ("ЭТО ФОРМУЛА", "В EXCEL!")
```

Мы *выделяем курсивом* новые термины и важные слова при первом упоминании. Названия файлов, веб-страниц и формул в тексте выглядят так:

<http://www.john-foreman.com>.

Итак, начнем

В первой главе я намерен заполнить некоторые пробелы в ваших познаниях об Excel, после чего вы сможете погрузиться непосредственно в практику. К концу книги вы не только будете иметь представление о нижеперечисленных техниках, но и приобретете опыт их применения:

- оптимизация с использованием линейного и интегрального программирования;
- работа с временными рядами данных, определение трендов и изменений сезонного характера, а также прогнозирование методом экспоненциального сглаживания;
- моделирование методом Монте-Карло в оптимизации и прогнозировании сценариев для количественного выражения и адресации рисков;
- искусственный интеллект с использованием общей линейной модели, функции логистических звеньев, ансамблевых методов и наивного байесовского классификатора;
- измерение расстояния между клиентами с помощью близости косинусов угла, создание K-ближайших граф, расчет модулярности и кластеризация клиентов;
- определение выбросов в одном измерении по методу Тьюки или в нескольких измерениях с помощью локальных факторов выброса;
- применение пакетов R для использования результатов работы других программистов при выполнении этих задач.

Если хотя бы что-то из вышесказанного звучит для вас воодушевляюще — продолжайте чтение! Если пугающе — то тоже продолжайте! Я торжественно обещаю разжевывать все как можно тщательнее.

Итак, без лишней суеты приступим!

1

Все, что вы жаждали знать об электронных таблицах, но боялись спросить

В этой книге я исхожу из того, что вы уже имеете некоторое представление об электронных таблицах и пользуетесь ими. Если же вы никогда не сталкивались с расчетами по формулам, вам поначалу придется нелегко. Перед нашим совместным погружением в Excel с головой я бы рекомендовал проштудировать «Excel для чайников» или другую подобную литературу вводного уровня.

Но даже если вы — заслуженный мастер по работе с Excel, все равно в моем тексте иногда будут возникать упоминания о таких возможностях программы, которыми вы никогда не пользовались. Так, в настоящей главе вы встретите много небольших приемов с простыми функциями. Некоторые наверняка покажутся вам немного странными. Не зацикливайтесь на непонятном — двигайтесь дальше. Вы всегда сможете вернуться к недочитанной главе позже.

ОТЛИЧИЯ В РАЗНЫХ ВЕРСИЯХ EXCEL

Как я уже упоминал во введении, для этой книги подходят Excel 2007, 2009, 2011 для MacOS и LibreOffice. К сожалению, в каждой новой версии Excel разработчики Microsoft перемещают инструменты и функционал как им угодно.

Например, элементы из вкладки «Разметка» в версии 2011 года находятся во вкладке «Вид» во всех остальных версиях. «Поиск решения» в версиях 2010 и 2013 одинаковый, но реализован он лучше в 2007 и 2011, несмотря на гротескный интерфейс в версии 2007.

Снимки с экрана в этой книге будут делаться с Excel 2011. Если у вас более новая или старая версия, вам придется действовать по-другому, особенно когда дело касается положения элементов управления во вкладках меню. Я очень постараюсь найти и учесть все различия. Если что-то мной упущено, поисковый инструмент Excel и Google всегда к вашим услугам.

А вот то, что, несомненно, должно вас обрадовать: «табличная часть электронной таблицы» всегда неизменна.

Несколько слов о LibreOffice. Если вы решили пользоваться открытыми источниками программного обеспечения, рискну предположить, что вы – человек, склонный до всего доходить самостоятельно. И хотя я не буду напрямую обращаться к интерфейсу LibreOffice, вы этого попросту не заметите. Они с Excel похожи как две капли воды.

Немного данных для примера

ЗАМЕТКА:

Рабочая тетрадь Excel, используемая в этой главе «Concessions.xlsx», доступна для загрузки на сайте книги www.wiley.com/go/datasmart

Представьте себе, что вам жутко не везет по жизни. Даже став взрослым, вы до сих пор живете с родителями и работаете в киоске на баскетбольных матчах в своей старой школе. (Клянусь, это только наполовину автобиографично!).

У вас есть электронная таблица о вчерашних продажах, и выглядит она примерно как рис. 1-1.

	A	B	C	D
1	Item	Category	Price	Profit
2	Beer	Beverages	\$ 4.00	50%
3	Hamburger	Hot Food	\$ 3.00	67%
4	Popcorn	Hot Food	\$ 5.00	80%
5	Pizza	Hot Food	\$ 2.00	25%
6	Bottled Water	Beverages	\$ 3.00	83%
7	Hot Dog	Hot Food	\$ 1.50	67%
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%
9	Soda	Beverages	\$ 2.50	80%
10	Chocolate Bar	Candy	\$ 2.00	75%
11	Hamburger	Hot Food	\$ 3.00	67%
12	Beer	Beverages	\$ 4.00	50%
13	Hot Dog	Hot Food	\$ 1.50	67%
14	Licorice Rope	Candy	\$ 2.00	50%
15	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%

Рис. 1-1. Продажи киоска

На рис. 1-1 показана каждая продажа: что именно продано, к какому типу еды или напитков относится проданный товар, цена и процент прибыли с продажи.

Быстрый просмотр с помощью кнопок управления

Если хотите ознакомиться с записями – промотайте список колесиком мышки, пальцем (если у вас сенсорный экран) или стрелками клавиатуры. Пока вы просматриваете записи, приятно иметь строку заголовков в поле зрения – тогда вы точно не забудете, что в какой колонке записано. Для этого выберите «Закрепить области» или «Закрепить верхнюю строку» во вкладке «Вид» в Windows (вкладка «Разметка» в MacOS2011, как показано на рис. 1-2).

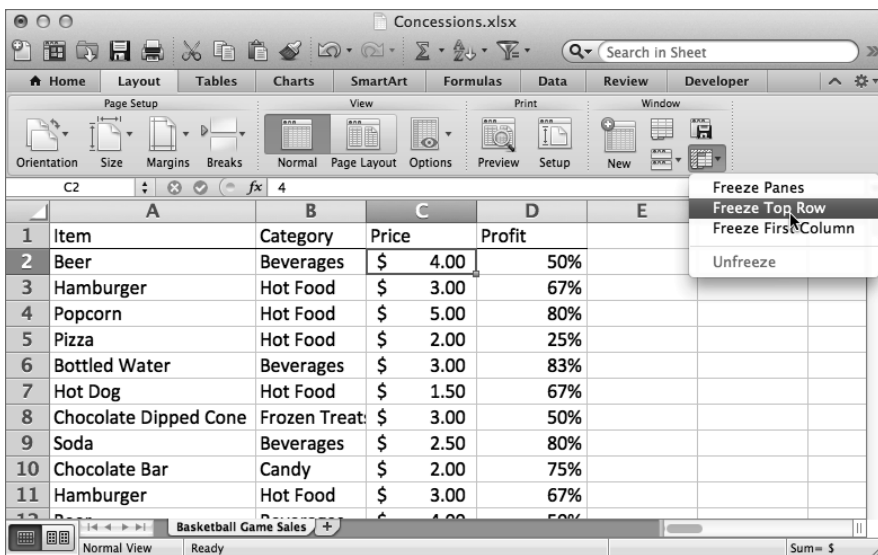


Рис. 1-2. Закрепление верхней строки.

Чтобы быстро переместиться в конец документа и посмотреть, сколько всего у вас продаж, выберите значение в одном из заполненных столбцов и нажмите **Ctrl+↓** (**Command+↓** на Mac). Вас отбросит прямо к последней заполненной ячейке этого столбца. В этой таблице последняя строка – 200. Заметьте также, что кнопка **Ctrl/Command** со стрелками даст вам возможность точно так же перемещаться по всему документу, в том числе вправо и влево.

Если вы хотите узнать среднюю прибыль за единицу проданного за вечер, то под столбцом с ценами (столбцом C) можно вбить формулу:

```
=AVERAGE (C2 : C200) /
=CPЗНАЧ (C2 : C200)
```

Средняя прибыль получается \$2,83, так что отдыхать от трудов праведных вам, увы, еще не время. Подсчет можно произвести и другим способом: переместиться на последнюю ячейку в столбце, C200 и удерживать Shift+Ctrl+↑, чтобы выделить весь столбец доверху, а затем выбрать «Среднее значение» на нижней панели справа (рис. 1-3). В Windows нужно кликнуть на этой панели для того, чтобы выбрать среднее значение, скажем, вместо суммы, стоящей там по умолчанию. В MacOS, если нижняя панель отключена, нужно нажать на меню «Вид» и выбрать «Строку состояния», чтобы включить ее.

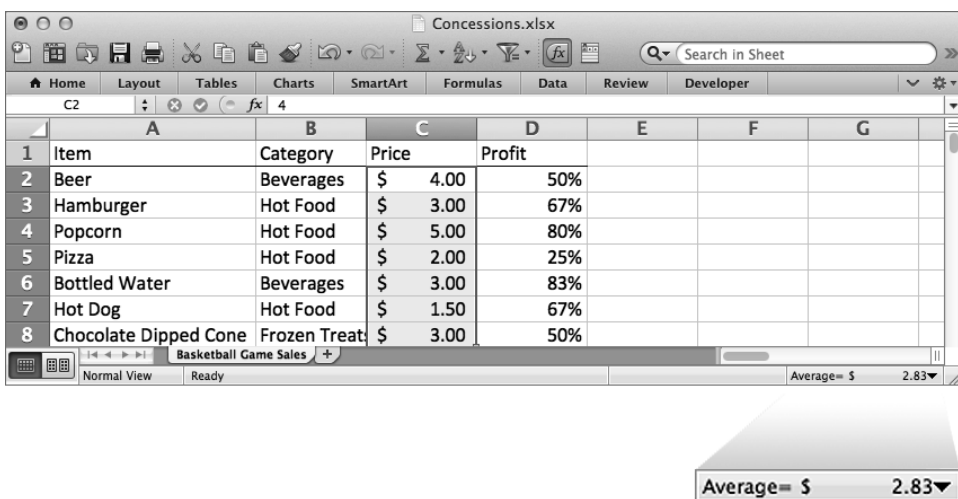


Рис. 1-3. Среднее значение колонки с ценами на панели состояния

Быстрое копирование формул и данных

Пожелав видеть свою прибыль в фактических долларах, а не в процентах, вы можете добавить что-то вроде заголовка в столбец E, который назовем «Фактическая прибыль». В ячейке E2 нужно просто перемножить соответствующие значения из столбцов с ценой и прибылью, чтобы получить такую формулу:

=C2*D2

Для строки с пивом результат будет равным \$2. Не нужно переписывать формулу для каждой строки. Excel позволяет переносить формулы из ячейки перетаскиванием за правый нижний угол куда вам угодно. Значения в столбцах C и D будут меняться в зависимости от того, куда скопирована формула. Если, как в случае с данными о продажах, столбец слева полностью заполнен,

1. Все, что вы жаждали знать об электронных таблицах, но боялись спросить

дважды кликните на правом нижнем углу ячейки с формулой, и Excel заполнит значениями весь столбец, как это показано на рис. 1-4. Попробуйте этот двойной клик сами, потому что я буду использовать его во всей книге. Освоив его сейчас, вы избавите себя от огромных неудобств в будущем.

Обязательно ли значение в ячейках, упомянутых в формуле, должно меняться в зависимости от того, куда вы ее перетаскиваете или копируете? Нет, конечно. Хотите оставить что-то неизменным – просто поставьте перед ним \$.

К примеру, если вы измените формулу в E2 таким образом:

=C\$2×D\$2

	A	B	C	D	E
1	Item	Category	Price	Profit	Actual Profit
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00
3	Hamburger	Hot Food	\$ 3.00	67%	
4	Popcorn	Hot Food	\$ 5.00	80%	
5	Pizza	Hot Food	\$ 2.00	25%	
6	Bottled Water	Beverages	\$ 3.00	83%	
7	Hot Dog	Hot Food	\$ 1.50	67%	
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	

Рис. 1-4. Заполнение столбца по формуле растягиванием ее за угол

В этом случае при копировании формулы на все последующие ячейки в ней ничего не меняется. Формула продолжает обращаться ко 2-й строке.

Если скопировать формулу вправо, то C заменится на D, D на E и т.д. Если вам не нравится такое «поведение», добавьте \$ также перед ссылками на столбцы в формуле. Это называется *абсолютной ссылкой*, в противоположность *относительной ссылке*.

Форматирование ячеек

Excel предлагает статические и динамические опции для форматирования содержимого ячеек. Взгляните на столбец E с фактической прибылью, который вы только что создали. Выделите его, кликнув на серый заголовок колонки. Затем кликните на выбранном столбце правой клавишей и выберите «Формат ячеек».

В этом меню вы можете выбрать формат содержимого ячеек столбца E. В нашем случае нужен денежный формат. Также можно указать число знаков после запятой при округлении. Оставьте 2 знака после запятой, как показано на рис. 1-5. Также в меню «Формат ячеек» доступны такие опции, как цвет шрифта, заливка ячейки, выравнивание текста, границы и т. д.

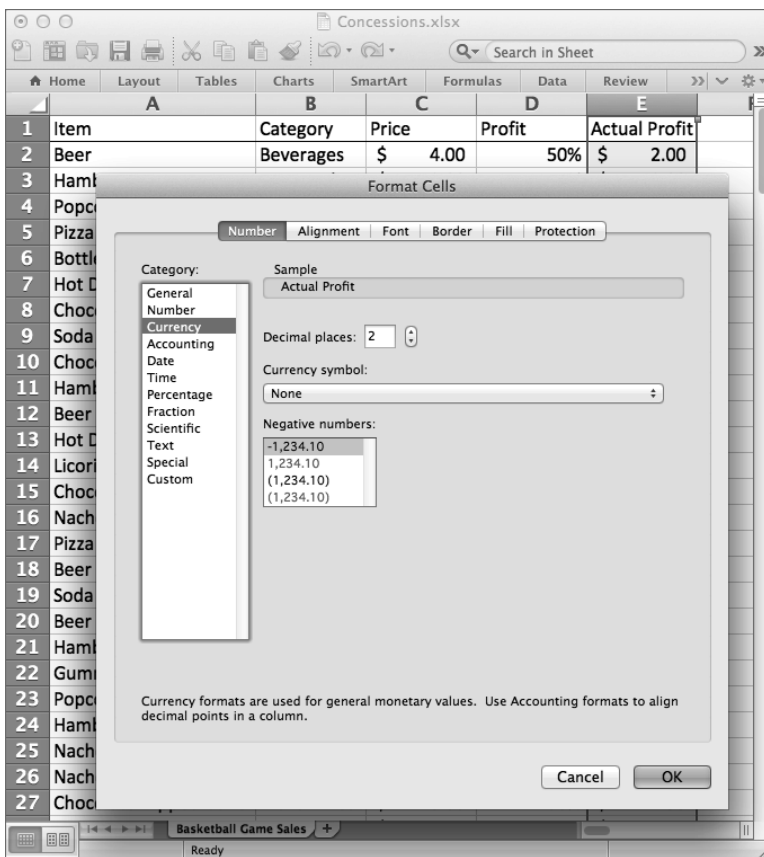


Рис. 1-5. Меню «Формат ячеек»

Но есть нюанс. Допустим, нужно отформатировать только те ячейки, которые содержат определенные значения или диапазон значений, и это форматирование должно меняться в зависимости от значений.

Такой вид форматирования называется условным форматированием, и оно повсеместно используется в этой книге.

Закройте меню «Формат ячеек» и переместитесь во вкладку «Главная». В разделе «Стили» («Формат» в MacOS) вы найдете «Условное форматирование»

(рис. 1-6). При нажатии на него выпадает меню. Самое используемое условное форматирование в этой книге — цветовые шкалы. Выберите шкалу для столбца E и посмотрите, как изменился цвет каждой ячейки в зависимости от величины значения в ней.

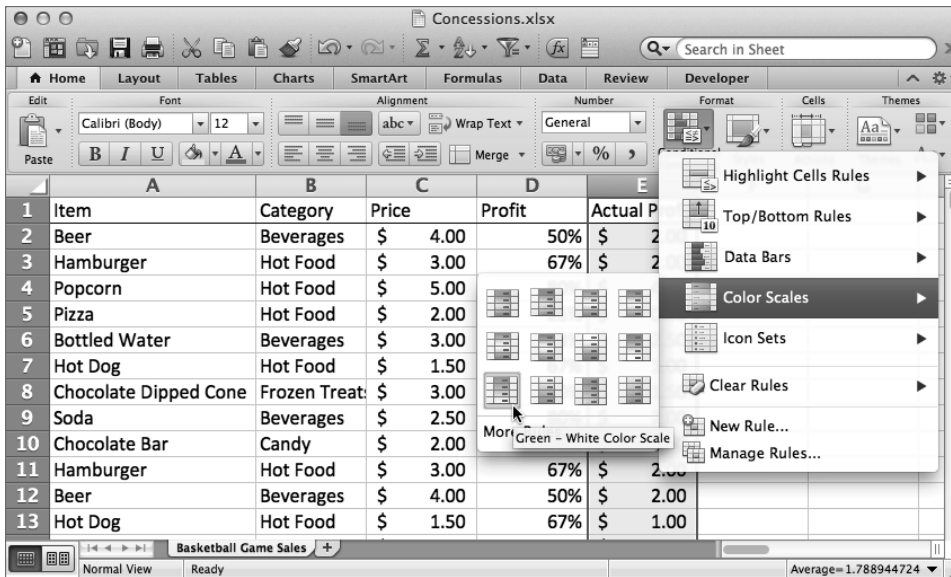


Рис. 1-6. Применение условного форматирования к столбцу с прибылью

Чтобы очистить условное форматирование, используйте опцию «Удалить правила» меню условного форматирования.

Специальная вставка

Конечно, гораздо удобнее работать, если формулы не путаются у вас под рукой, как в колонке E на рис. 1-4. А если это еще и формулы вроде `RAND()`/`СЛЧИС()`, генерирующей случайные числа, которые меняют свое значение при каждом автопересчете таблицы, то ваше раздражение вполне справедливо. Решение проблемы — в копировании этих ячеек и вставке их обратно в таблицу в виде постоянных величин.

Чтобы перевести формулы в цифры, просто выделите и скопируйте столбец E, заполненный формулами, и вставьте его обратно с помощью опции «Специальная вставка» (находится во вкладке «Главная» под опцией «Вставить» в Windows

и в меню «Редактировать» в MacOS). В окне «Специальная вставка» выберите вставку в качестве значений (рис. 1-7). Замечу, что это меню при вставке позволяет также транспонировать данные из вертикали в горизонталь и наоборот. Это свойство очень пригодится вам в дальнейшем.

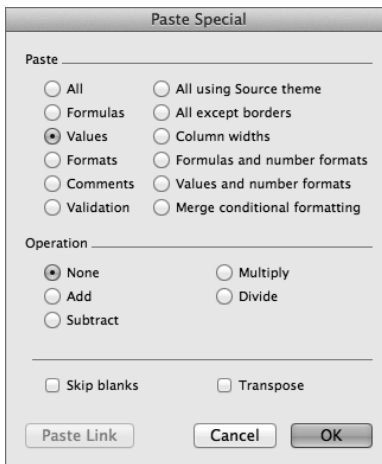


Рис. 1-7. Окно «Специальная вставка» в Excel 2011

Вставка диаграмм

Методичка, посвященная торговле с лотка, включает в себя графу «Калории» с малюсенькой табличкой. В ней указано, сколько калорий содержится в каждом напитке или закуске, которые продаются в киоске. Вы тоже легко можете сделать такую диаграмму в Excel. Во вкладке «Вставка» («Диаграммы» в MacOS) есть раздел, в котором находятся различные варианты отображения, такие как столбчатая гистограмма, линейный график и круговая диаграмма.

ЗАМЕТКА

В этой книге мы в основном будем пользоваться столбчатыми гистограммами, линейными графиками и графиками рассеяния. Никогда не пользуйтесь круговыми диаграммами! И особенно круговыми 3D-диаграммами, которые вам предлагает Excel. Не вздумайте послушаться, иначе мой призрак будет мучить вас после моей смерти! Круговые диаграммы уродливы, плохо соотносятся с данными, эстетически их 3D-эффект примерно таков же, как у картинок из ракушек на стене кабинета моего стоматолога.

1. Все, что вы жаждали знать об электронных таблицах, но боялись спросить

Выделяя столбцы А и В в листе «Calories» вы можете выбрать столбчатую диаграмму с группировкой для отображения данных. Поиграйте с графикой. Нажимайте на разделы правой клавишей мыши, чтобы увидеть меню форматирования. Например, щелчок правой клавиши на столбцах диаграммы вызовет меню, в котором можно выбрать «Формат рядов данных». Под ним вы сможете поменять цвет столбцов с синего по умолчанию на любой оттенок, который вам по вкусу, например черный.

В наличии легенды по умолчанию тоже нет никакого смысла, так что совету выделите ее и нажмите «Удалить». Также вам может понадобиться выделить разные текстовые подписи к диаграмме и увеличить размер шрифта (размер шрифта находится под вкладкой «Главная»). Таким образом получается диаграмма, показанная на рис. 1-8.

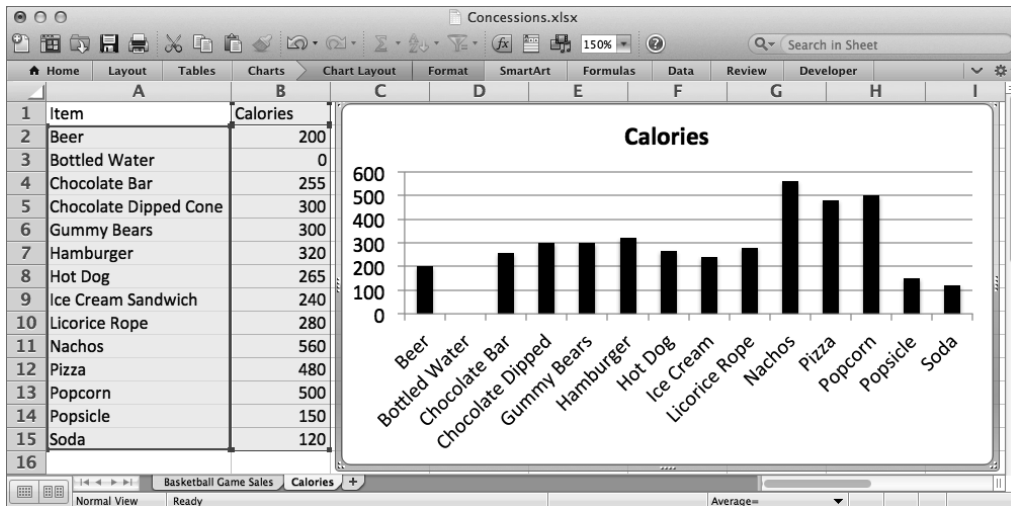


Рис. 1-8. Вставка диаграммы содержания калорий

Расположение меню поиска и замены

В этой книге вам частенько придется пользоваться функциями поиска и замены. В Windows это делается, как обычно, нажатием Ctrl+F для открытия окна поиска (и Ctrl+N для замены) или перемещением во вкладку «Главная», где в разделе «Правка» находится кнопка «Найти». В MacOS строка поиска расположена в верхнем правом углу листа (для замены нажмите либо стрелку вниз, либо Cmd+F для вызова меню поиска и замены).

Чтобы проверить прочитанное на практике, откройте меню замены на листе «Calories». Замените слово «Калории» на слово «Энергия» везде, где оно встречается (рис. 1-9), вбив эти слова в окно поиска и замены и нажав «Заменить все».

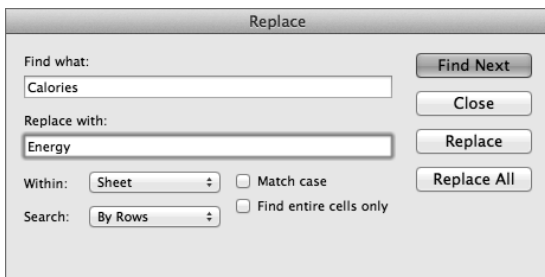


Рис. 1-9. Поиск и замена

Формулы поиска и вывода величины

Если бы я не уточнил, что вам знакомы хотя бы некоторые простые формулы Excel (SUM, MAX, MIN, PERCENTILE / СУММ, МАКС, МИН, ПЕРСЕНТИЛЬ и т. д.), мы бы просидели здесь целый день. А я хочу начать анализировать данные. Вместе с тем я часто использую в этой книге формулы, с которыми вы могли ни разу не столкнуться, если до этого не погружались с головой в волшебный мир электронных таблиц. Эти формулы работают с поиском значения в ряду и выводом его положения или, наоборот, поиском положения в ряду и возвратом значения.

Я покажу это на примере листа «Calories».

Иногда хочется узнать положение элемента в столбце или строке. Первый он, второй или третий? Формула МАТЧ/ПОИСКПОЗ справляется с этим довольно неплохо. Под вашими данными о калориях назовите A18 Match/Поискпоз. Вы можете применить формулу к ячейке B18, чтобы найти, где в списке выше упоминается слово «Hamburger». Чтобы использовать эту формулу, необходимо указать в ней значение, которое нужно найти, границы поиска и 0, чтобы она вывела позицию самого слова:

```
=МАТЧ("Hamburger", A2:A15, 0) /  
=ПОИСКПОЗ("Hamburger", A2:A15, 0)
```

Она выдает 6, так как «Hamburger» – шестая позиция в списке (рис. 1-10).

	A	B	C
12	Pizza	480	
13	Popcorn	500	
14	Popsicle	150	
15	Soda	120	
16			
17			
18	Match	6	
19	Index	0	
20	Offset	Chocolate Bar	
21	Small	150	

Рис. 1-10. Формулы, которые нужно знать

Следующая формула — INDEX / ИНДЕКС. Назовите ячейку A19 Index/Индекс.

Эта формула находит значение элемента по заданному положению в строке или столбце. Например, подставив в нее из нашей таблицы калорий A1:B15 и задав координаты поиска «3 строка, 2 столбец», мы получим количество калорий в бутылке воды:

```
=INDEX (A1 : B15, 3, 2) /
=ИНДЕКС (A1 : B15, 3, 2)
```

Мы видим количество калорий, равное 0, как и предполагалось (рис. 1-10).

Другая формула, которая часто встречается в нашем тексте, — это OFFSET / СМЕЩ. Назовем же ячейку A20 Offset/Смещ и поиграем с формулой в B20.

С помощью этой формулы вы задаете промежуток, который перемещаете, подобно курсору, по сетке из столбцов и строк (точно так же, как INDEX/ИНДЕКС ищет единственную ячейку, если только в нем не упомянут 0). Например, можно задать функции OFFSET/СМЕЩ рамки от верхней левой ячейки листа A1 и затем растянуть ее на 3 ячейки вниз, создавая ряд из 3 строк и 0 столбцов:

```
=OFFSET (A1, 3, 0) /
=СМЕЩ (A1, 3, 0)
```

Эта формула возвращает значение третьего элемента списка — «Chocolate Bar» (рис. 1-10).

Последняя формула, о которой я хочу сказать в этом разделе, — SMALL/НАИМЕНЬШИЙ (у него есть двойник — LARGE/НАИБОЛЬШИЙ, который работает точно

так же). Если у вас есть список значений и вы хотите выбрать, скажем, третье наименьшее из них, данная функция делает это за вас. Назовите ячейку A21 Small/Наименьший, а в B21 напишите следующую формулу, содержащую границы поиска и параметр 3:

```
=SMALL (B2 : B15 , 3) /  
=НАИМЕНЬШИЙ (B2 : B15 , 3)
```

Эта формула возвращает значение 150, которое является третьим наименьшим после 0 (бутылка воды) и 120 (газировка), как показано на рис. 1-10.

И, наконец, еще одна формула для поиска значений, похожая на МАТЧН/ПОИСКПОЗ, употребившую стероиды. Это VLOOKUP/ВПР (и ее горизонтальный двойник HLOOKUP/ГПР). Им я уделю целый раздел, ибо это монстры.

Использование VLOOKUP/ВПР для объединения данных

Перейдем обратно к листу продаж на баскетбольных матчах. При этом мы в любое время можем обратиться предыдущему листу с калориями, просто указав его название и поставив перед номером ячейки «!». Например, Calories!B2 является ссылкой к количеству калорий в пиве, несмотря на то, что вы в данный момент работаете с другим листом.

Предположим, вы захотите увидеть количество калорий на листе продаж для каждого наименования товара. Вам нужно будет каким-то образом найти содержание калорий в каждом товаре и поместить его в колонку, следующую за прибылью. Что ж, оказывается, и для этого есть отдельная функция под названием VLOOKUP/ВПР.

Назовем колонку F в нашем листе «Calories / Калории». Ячейка F2 будет содержать количество калорий из таблицы в товаре из первой строки — пиве. Используя эту формулу, можно указать в названии товара из ячейки A2 ссылку на таблицу Calories!\$A\$1:\$B\$15 и номер столбца, из которого следует выбирать значения. В нашем случае он второй по счету:

```
=VLOOKUP (A2, Calories!$A$1 : $B$15 , 2 , FALSE) /  
=ВПР (A2, Calories!$A$1 : $B$15 , 2 , ЛОЖЬ)
```

FALSE/ЛОЖЬ в конце формулы означает, что вам не подходят приблизительные значения «Beer». Если функция не может найти «Beer» в таблице калорий, она возвращает ошибку.

1. Все, что вы жаждали знать об электронных таблицах, но боялись спросить

После ввода формулы вы увидите, что 200 калорий считались из таблицы в листе «Calories». Поставив \$ в формуле перед ссылками на таблицу, вы можете скопировать формулу вниз по колонке двойным щелчком на нижнем правом углу ячейки. Оп-ля! У вас есть количество калорий для каждой позиции, как показано на рис. 1-11.

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00	200
3	Hamburger	Hot Food	\$ 3.00	67%	\$ 2.00	320
4	Popcorn	Hot Food	\$ 5.00	80%	\$ 4.00	500
5	Pizza	Hot Food	\$ 2.00	25%	\$ 0.50	480
6	Bottled Water	Beverages	\$ 3.00	83%	\$ 2.50	0

Рис. 1-11. Использование VLOOKUP/ВПР для указания количества калорий

Фильтрация и сортировка

Отразив в листе продаж калорийность ваших товаров, задайтесь целью видеть, например, только товары из категории «Замороженные продукты» — иными словами, отфильтровать ваш лист. Для этого сначала выберите данные в рамках A1:F200. Наведите курсор на A1 и нажмите Shift+Ctrl+↓, а затем →. Есть способ еще проще — кликнуть наверху столбца и, удерживая клавишу мышки нажатой, переместить курсор к столбцу F, чтобы выделить все 6 столбцов.

Затем, чтобы применить автофильтрацию к этим шести колонкам, нажмите кнопку «Фильтр» из вкладки «Данные». Она похожа на серую воронку, как на рис. 1-12.

Если автофильтрация включена, можно кликнуть на выпадающем меню, которое появляется в ячейке B1, и выбрать для показа только определенные категории (в данном случае отобразятся товары из категории «Замороженные продукты»), как на рис. 1-13.

После фильтрации выделение столбцов данных позволяет нижней панели показывать краткую информацию об этих ячейках. Например, отфильтровав только замороженные продукты, можно выделить значения в столбце E и использовать нижнюю панель, чтобы быстро узнать сумму прибыли только по этой категории товара, как на рис. 1-14.

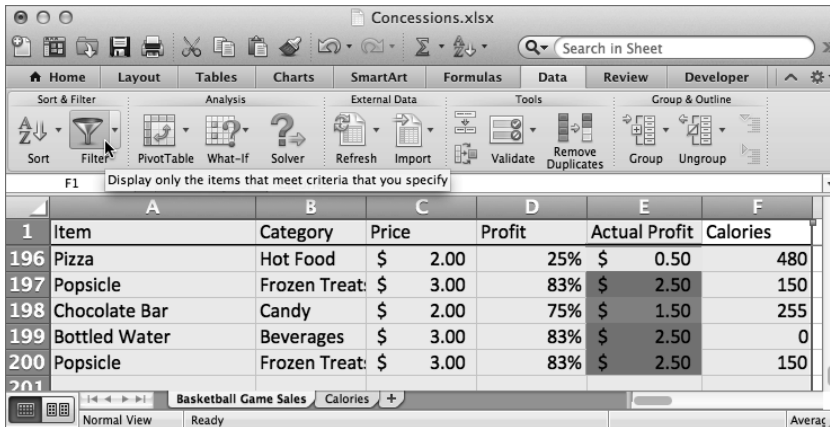


Рис. 1-12. Автофильтрация выбранной области

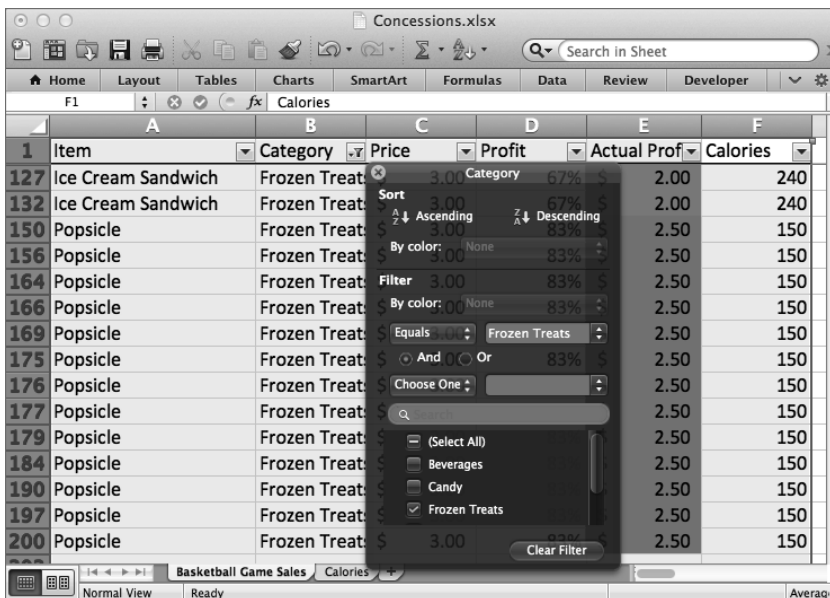


Рис. 1-13. Фильтрация по категории

Автофильтрация позволяет также производить сортировку. К примеру, если вы хотите рассортировать прибыль, просто кликните на меню автофильтрации в ячейке Profit/Прибыль (D1) и выберите сортировку по возрастанию (или убыванию), как на рис. 1-15.