Walmart Data Exploration

● ● ●

Questions & Answers on Anonymized Data

@ramonsuarez   ramonsuarez.com   linkedin.com/in/ramonsuarez

Table of Contents

- Why I chose this dataset
- About Walmart
- The dataset
- Assumptions
- Sales introduction
- Types of stores
- Seasonality
- Markdowns
- Temperature
- Challenges & what I've learned
- To improve
- How I worked
- What's next
- Q&A

Why I chose this dataset

- Business data:
  - Pain points in career
  - Use knowledge to advance business
- Time series:
  - Search for trends, seasonality, holiday & promos impact
  - Forecasting

About Walmart

- Funded in 1962 by Sam Walton
- 10,586 stores
- 24 countries
- $570 billion annual revenue
- 2.2 million employees

The dataset

- 45 USA stores
- 3 types of store
- 400K+ rows
- Feb 5 2010 to Nov 1 2012*

kaggle

Markdown data is only available after Nov 2011, and is not available for all stores all the time

Assumptions

- The initial SQL DB will emulate a production DB, so the import of the original files will only have to be made once.
- Negative sales mean that there have been more returns than sales
- Store size is in square feet
- Temperature is in degrees Fahrenheit
- Money is US$
- Long numbers in Unemployment and CPI are decimal numbers missing the separator
- US formatting of dates and numbers
- Department numbers refer to the same departments across all stores

Sales Intro

- Total sales in period: $606,10 bn
- Best week in store: $69,31 m
- Worst week in store: $-498,89 k

Types of stores

- Types represent income buckets
- Can't tell if same as official types of stores because of size similarities

Seasonality

- Thanksgiving & Christmas bring the more revenue
  - Much smaller peak in lower income stores
- Superbowl & Saint Valentine weeks bring stores out of January slump

Sales vs Markdowns

- Markdowns seem to be correlated with increases in sales

Sales vs Temperature

- False correlation between sales and temperature driven by end of year sales?
- Most stores in dataset are not in north of US

Annual Average Max Temperature (°F)

http://climatemaps.co.cc
Source: NOAA NCDC 1961-1990
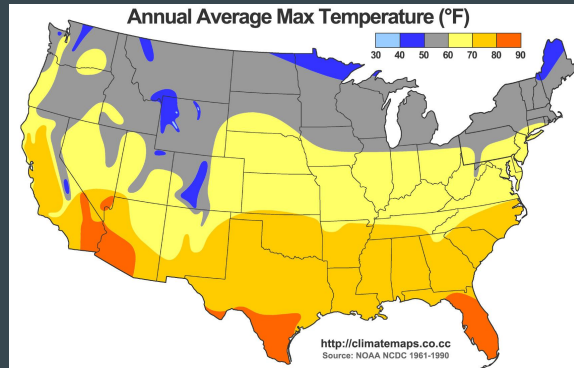
Challenges & what I've learned

- Challenges:
  - Not enough knowledge of statistics
  - Could not do forecasting
  - Remote work issues with SSIS & Power BI
  - Long delays for minor issues
  - Working with weekly sales data
  - Highlight chart areas
  - Renaming booleans on chart
- Learned:
  - Improved data analytics knowledge
  - Improved practical knowledge of tools
  - Documented issues and improved intuition
  - I want to do a lot more in data analytics

To improve

- Statistics
- Forecasting
- Machine learning

How I worked

- Toolstack:
  - TSQL & SSMS
  - Visual Studio SSIS
  - Power BI
  - Google Slides
  - Obsidian + Hugo

What's next

- Machine learning & statistics courses
- Data exploration with Python
- Power BI certification

Q&A

Questions?