‹XXXX›

Probation task for BI analyst position

Performed by Ilya Smolenskiy

XX/XX/XXXX

Steps in completing the probation task ‹XXXX›

- ✓ **Choose your approach:** *«Data Discovery»*
- ✓ **Choose a tool:** *Jupyter notebook > Power BI*
- ✓ **Choose a dataset:** *«Sample - Superstore.xls» from Tableau Community*
- ✓ **Do the analysis:**
  1. *Data preparation and exploration in Jupyter notebook*
  2. *Data visualization in Power BI Report*
- ✓ **Make a presentation about your findings:** *Here it is*
- ✓ **Send us the results in email:** *If you're reading this – you've received it*
- ○ **Impress us!** *- TBD*

Action plan

‹XXXX›

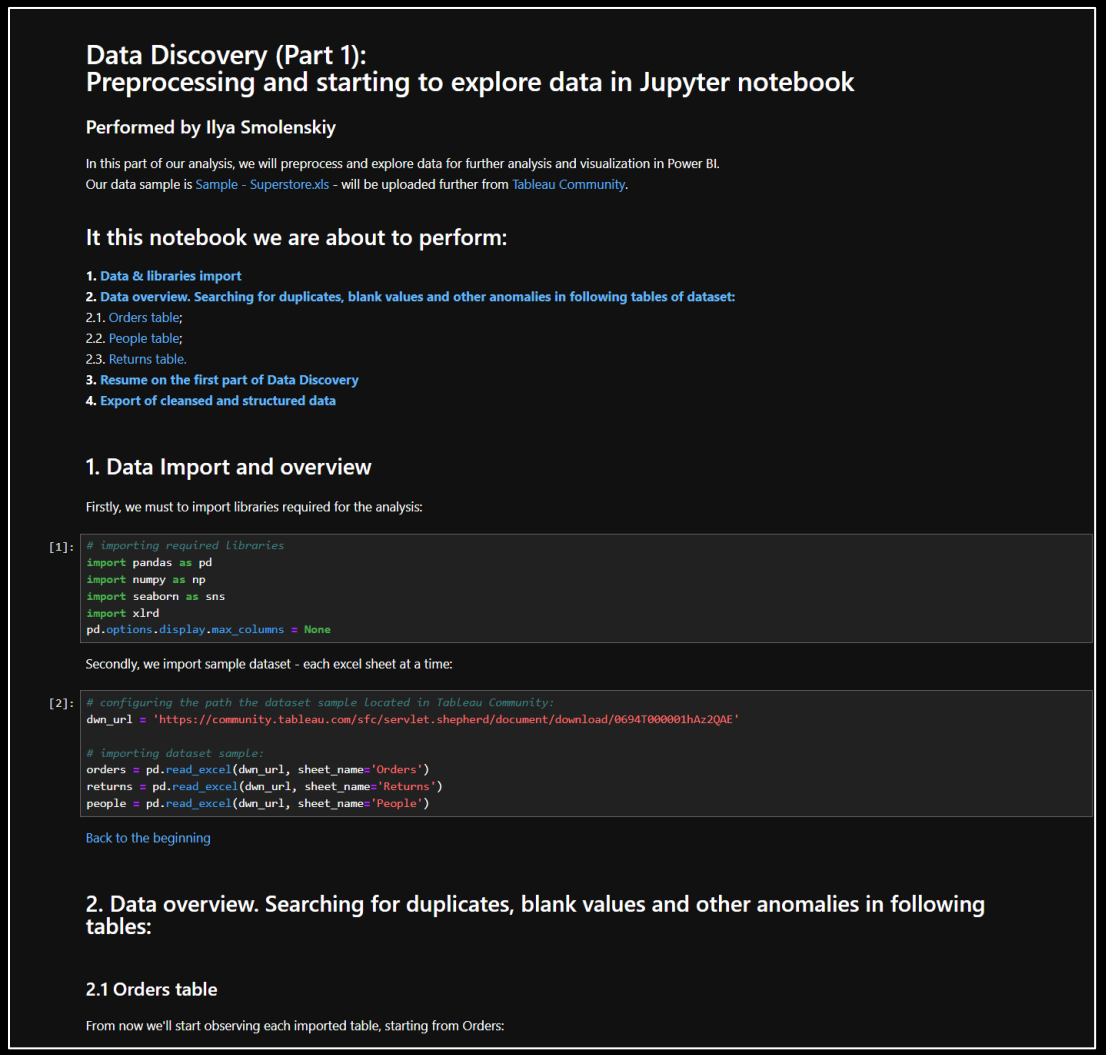
1. **Part 1 of Data Discovery:**
   Preprocessing and starting to explore data in Jupyter notebook:
   - Import & overview
   - Preprocessing and cleaning: elimination of duplicates, blank values etc.
   - Discovering interconnections and structures. Exploring behavior of variables
   - Searching for anomalies
   - Transformation and export
2. **Part 2 of Data Discovery:**
   Exploration and visualization of data in Power BI:
   - Query and preparation
   - Data model assembly
   - Calculation of additional variables: measures, columns, tables
   - Creating 3 tabs: Overview, Regional, Product
   - Filling tabs with comprehensive visuals
3. **Preparing a consolidated presentation with results and findings**

Part 1 of Data Discovery: Preprocessing and starting to explore data in Jupyter notebook. Overall Approach

‹XXXX›

During this part we've adopted **ETL** approach:

- **Extracted** data form Tableau Community straight to the notebook: Sample - Superstore.xls
- **Transformed**, as well as preprocessed and started to explore it utilizing Python
- **Loaded** it as separate tables into Google drive for further analysis in Power BI:
  - listofgoods.csv
  - cutomers.csv
  - ordercredentials.csv
  - orders.csv
  - people.csv
  - returns.csv



Data Discovery (Part 1): Preprocessing and starting to explore data in Jupyter notebook

Performed by Ilya Smolenskiy

In this part of our analysis, we will preprocess and explore data for further analysis and visualization in Power BI.
Our data sample is Sample - Superstore.xls - will be uploaded further from Tableau Community.

It this notebook we are about to perform:

1. Data & libraries import
2. Data overview. Searching for duplicates, blank values and other anomalies in following tables of dataset:
2.1. Orders table;
2.2. People table;
2.3. Returns table.
3. Resume on the first part of Data Discovery
4. Export of cleansed and structured data

1. Data Import and overview

Firstly, we must to import libraries required for the analysis:

[1]:
```
# importing required libraries
import pandas as pd
import numpy as np
import seaborn as sns
import xlrd
pd.options.display.max_columns = None
```

Secondly, we import sample dataset - each excel sheet at a time:

[2]:
```
# configuring the path the dataset sample located in Tableau Community:
dwn_url = 'https://community.tableau.com/sfc/servlet.shepherd/document/download/0694T000001hAz2QAE'

# importing dataset sample:
orders = pd.read_excel(dwn_url, sheet_name='Orders')
returns = pd.read_excel(dwn_url, sheet_name='Returns')
people = pd.read_excel(dwn_url, sheet_name='People')
```

Back to the beginning

2. Data overview. Searching for duplicates, blank values and other anomalies in following tables:

2.1 Orders table

From now we'll start observing each imported table, starting from Orders:

Part 1 of Data Discovery: Preprocessing and starting to explore data in Jupyter notebook. Dataset's description

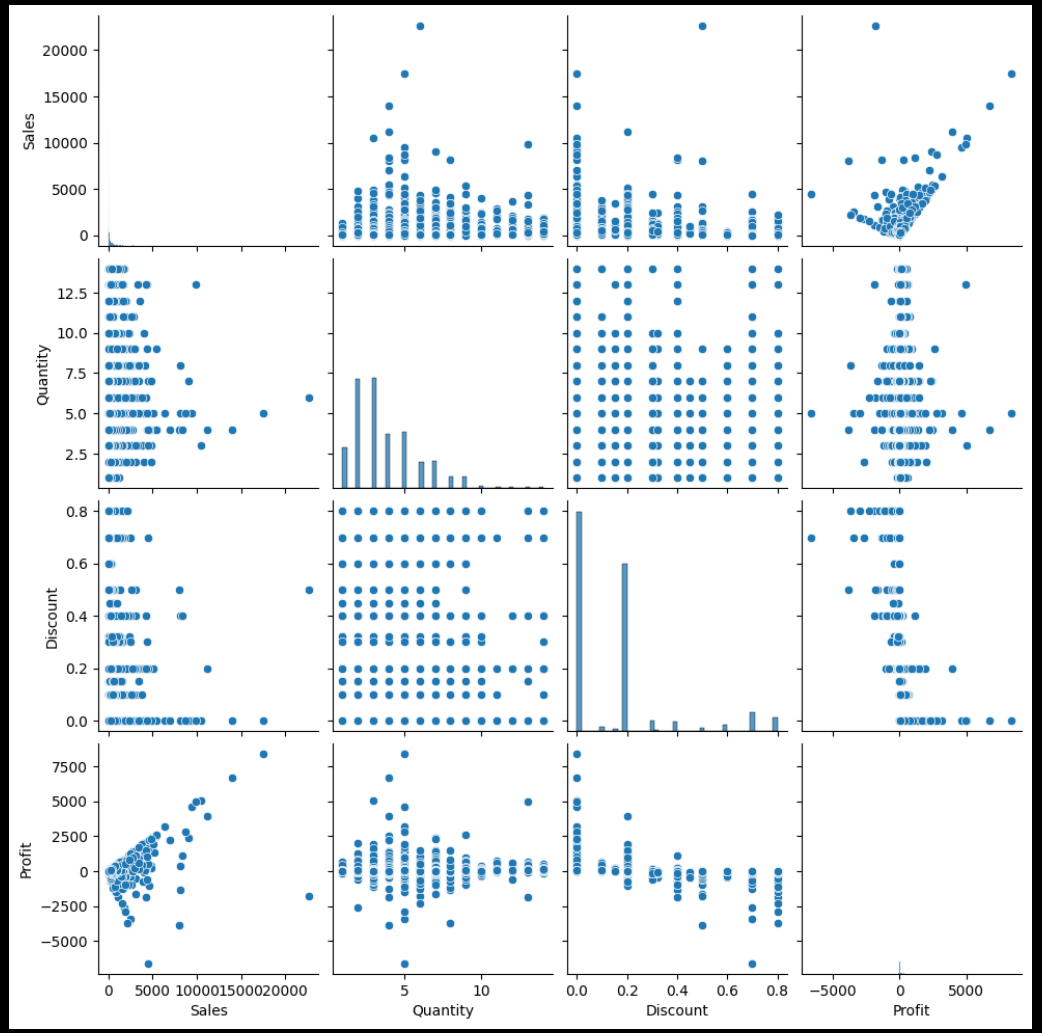
‹XXXX›

- Dataset must be an upload of an Appliance store data
- This store serves consumer and corporate customers in 49 US States
- This store sells products in categories: «Furniture», «Office Supplies» and «Technology»
- The dataset contains data of store's orders from 2011-01-04 to 2014-12-31 (~1458 days in total)
- Data set contained 3 related tables:
  - orders – each record in a table is a single item of an order, indicating quantity, sales, discount and profit, name, category, etc. per an item and additional information on shipping location and features, customer, etc. for each order.
  - people – a set of combinations of a regional manager name and region
  - returns – a set with indication on whether an order was returned for each order ID

Part 1 of Data Discovery: Preprocessing and starting to explore data in Jupyter notebook. Some finings in numerical variables:

‹XXXX›

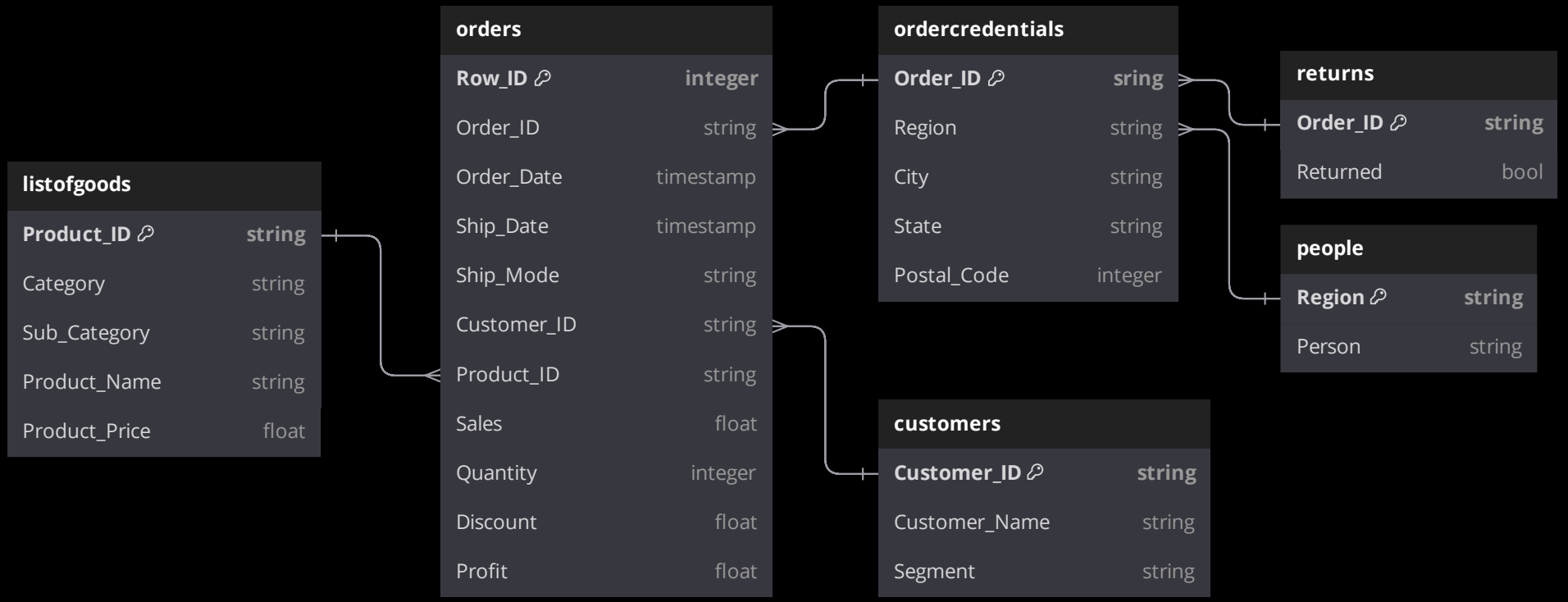
- Often users buy about 2-3 units of a single item
- Most losses of our profit are in zones of low bill sales in range between 0 and 5000
- The higher is discount the more chance to face a profit loss
- Highest discounts appear in the lower sales amount range
- Most of the times there are no discounts (zero discount) for a position in order



Part 1 of Data Discovery: Preprocessing and starting to explore data in Jupyter notebook. Data structure

‹XXXX›

Data structure examination unveiled MECE relations between variables, that leaded us to transforming the data model:



In this form of 6 related tables data was uploaded to Google Drive for further work in Power BI

Part 1 of Data Discovery: Preprocessing and starting to explore data in Jupyter notebook. Links to Python notebook

‹XXXX›

The notebook is also stored in Google Drive together with data outputs:

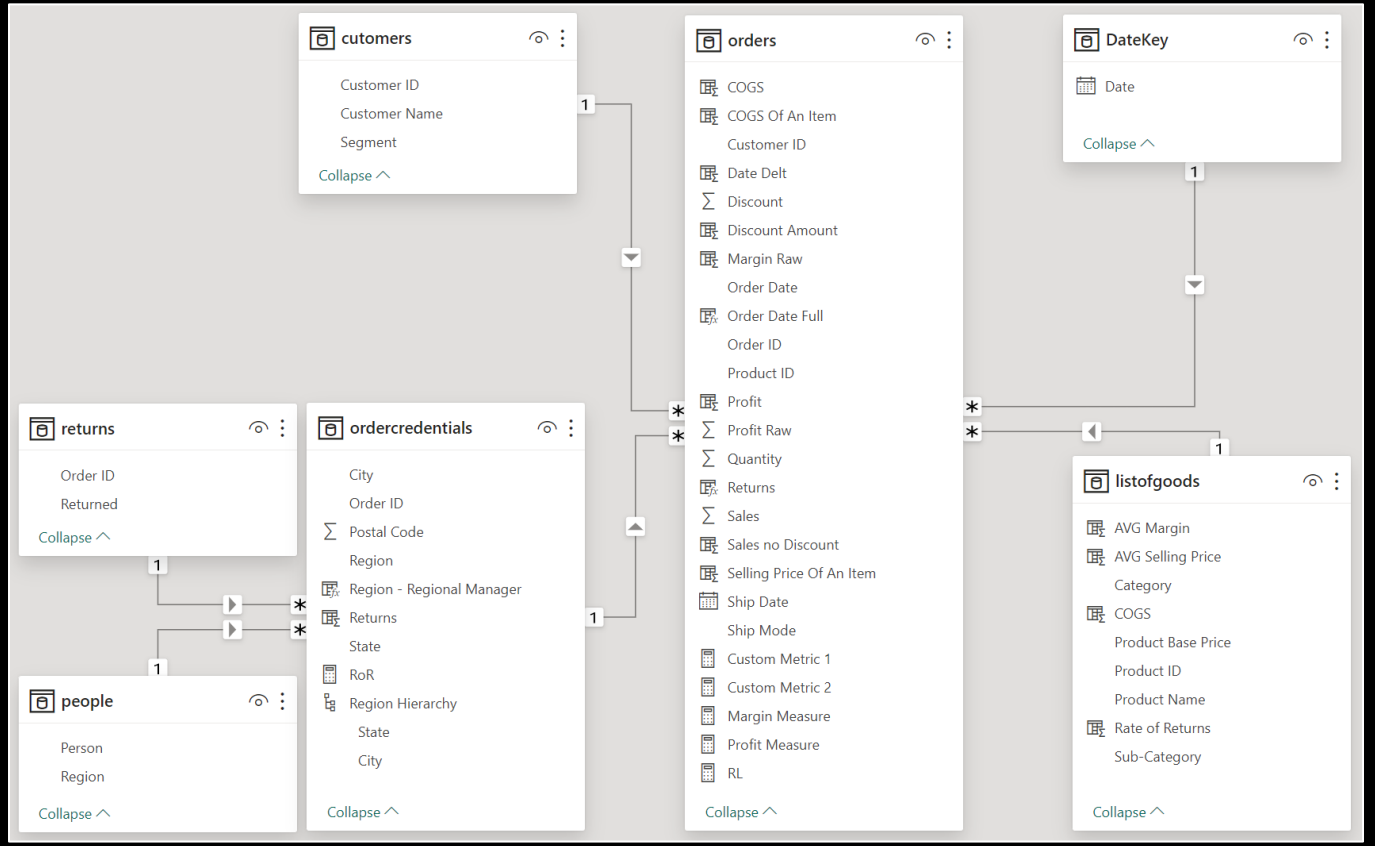
- Python notebook itself:
  01_Data Discovery (Part 1) Preprocessing and starting to explore data in Jupyter notebook.ipynb
- Notebook in HTML markdown (with code) – **download it to view in browser**:
  01_(with code) Data Discovery (Part 1) Preprocessing and starting to explore data in Jupyter notebook.ipynb
- Notebook in HTML markdown (no code) – **download it to view in browser**:
  01_(no code) Data Discovery (Part 1) Preprocessing and starting to explore data in Jupyter notebook.ipynb

Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Data model assembly

‹XXXX›

We've performed the following:

- Uploaded data directly from Google drive with almost no transformations of data in Power BI as we have already preprocessed it in Jupyter
- Assembled the data model in a form of star schema
- Set up various additional supporting metrics, based on given numeric variables incl.:
  - COGC - cost of goods sold
  - Rate of returns
  - Product and Profit Margin
  - etc.



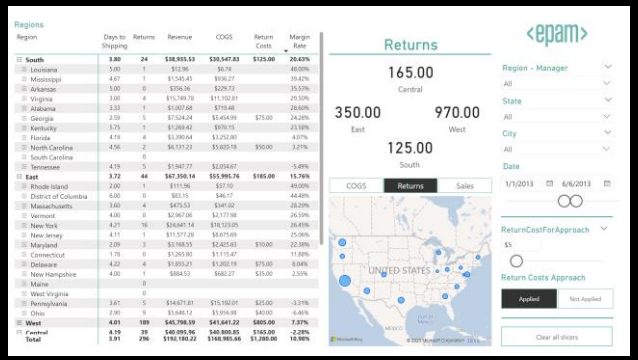
Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Dashboard structure

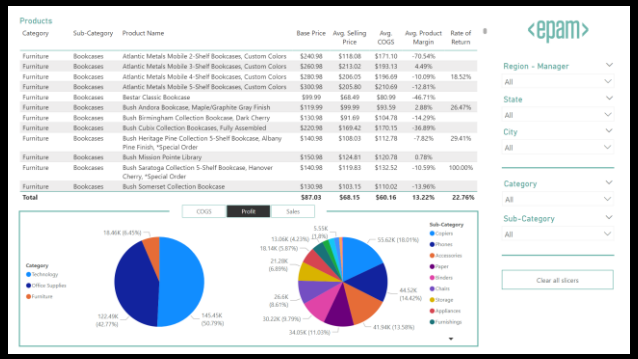
‹XXXX›

We've broken down our data to 3 tabs – each for single standpoints of observation:

- **Overview** – financial overview from historical and aggregate perspectives, mostly oriented on C-level stakeholders monitor of the key indices
- **Regional** – with a breakdown by states and cities this tab is designed to monitor expansion, control user experience related metrics (days to shipping, returns), indicate losses and extra profits with geographical reference.
- **Product** – mostly oriented on product managers to control sales performance, pricing and discounting policies for products / categories / subcategories

We've also provided each tab with a set of synchronized common and individual slicers for easy filtering.



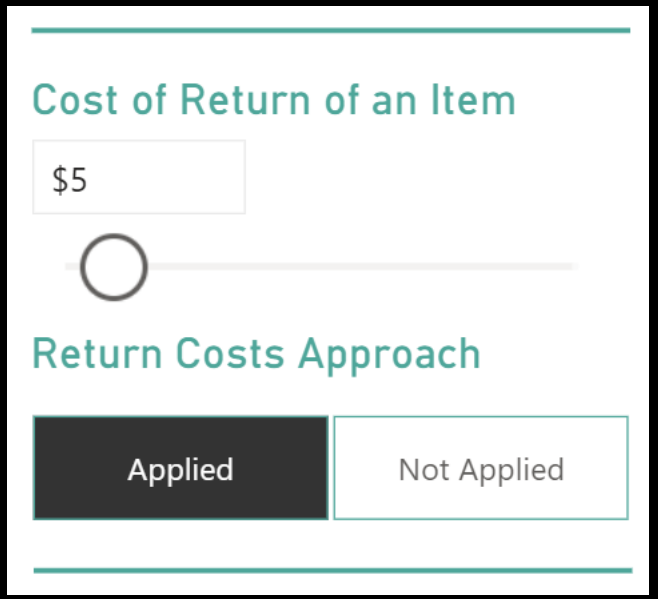


Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Return loss approach

‹XXXX›

To make this task more challenging we've implemented a Return loss approach:

- I case an order is returned, as we fully return money for the order to the customer, our Sales, COGS, Discount amount becomes None for this order.
- However, an additional cost of return of order's items back to the warehouse appears, so the profit for a returned order will be:



Profit = Quantity * Cost of Return of an Item

A report's user may define the Cost of Return amount or deactivate this approach by a set of slicers displayed on the right. It is activated by default with $5 Cost of Return

Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Findings in data (1/2)

‹XXXX›

- Observing store’s overall financial performance during the whole time frame we may say that it’s profitable with ~$ 250k in gross profit and a margin of ~10%.
- However, this profit margin seems low for an Appliance store business, comparing to alternative ways inverting capital to financial derivatives etc.
- Two most profit generating Regions are East (~$ 84k) and West (~$ 79k), supervised by Chuck Magee and Anna Andreadi regional managers respectively.
- However, these two regions are also leaning in the number of returns, generating additional costs of ~$3k and ~$10k correspondingly
- Texas with -$24k loss in Central region together with Pennsylvania with -$15k in East region are the most loss-making States of our store with profit margin less than -13% all time
- During the time of observation our store has expanded with clients from Wyoming, West Virginia, North Dakota and District of Columbia.
- The Columbia District yields the highest margin in the whole East region: ~37%, - whereas North Dakota generates up to 25% of store’s profit margin.

Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Findings in data (2/2)

‹XXXX›

- Goods from categories «Office Supplies» & «Technology» generate up to 93% of store's profit, however, in terms of sales previously mentioned categories' share is distributed equally with the third category of goods – «Furniture».
- Breaking down probits by sub-categories of goods, there are three generating up to 47% of all profits: «Copiers», «Phones», «Accessories».
- However, the highest rate of order return is in «Machines» ~58% and «Copiers» ~29%, representing the category of «Technology»
- The least average Days to Shipping metric is in the East region, leaded by West Virginia with 3 days and Rhode Island with 3.3 days – we may assume that our store has located an additional stock hub in between these states.

All the findings above are the starting points for deeper product- and region-oriented analysis to conduct loss prevention & profit maximization by improving store's list of products, logistics, customers' experience, etc.

Part 2 of Data Discovery: Exploration and visualization of data in Power BI. Links to Power BI report

‹XXXX›

The Report is stored in Google Drive:

- Interactive report in .pbix:
  02_Data Discovery (Part 2)
  Exploration and visualization of data in Power BI.pbix

- An exported preview in .pdf:
  02_Data Discovery (Part 2)
  Exploration and visualization of data in Power BI.pdf

Resume/Conclusion

‹XXXX›

During this probation we've performed advanced analysis including:

- **Extract – Transform – Load**, preprocessing and exploring data in Jupyter notebook
- **Interactive dashboard design**, exploring and visualizing data in Power BI
- **Summary presentation preparation**, overviewing task's outputs and key findings

The whole probation task's output is stored on my Google Drive

Thank you for your attention!