Big Data Analytics Project
-Part I-

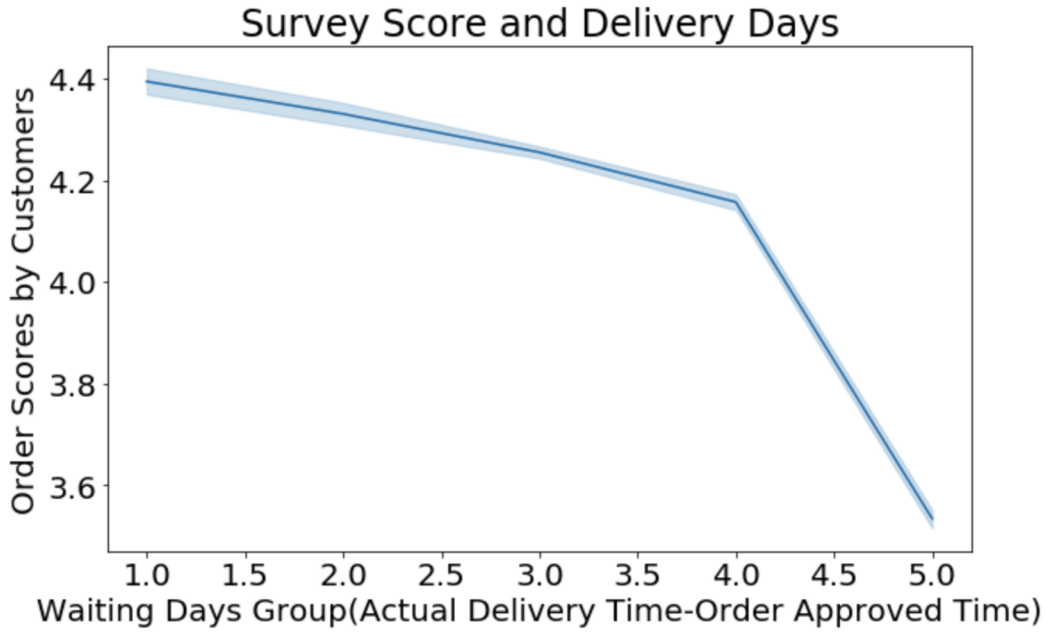
- **Dataset Selection:** Brazilian E-commerce Company Olist from Kaggle
- **Project Topic:** Analysis of the E-commerce Business of Olist on Optimizing Logistics Solutions and Increasing Customer Experiences

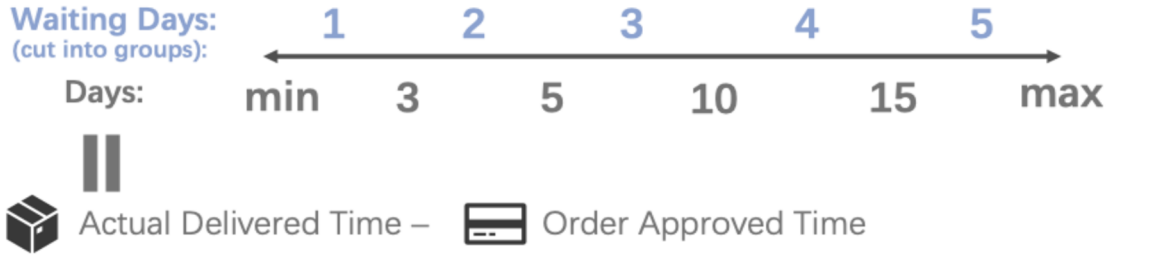
Group Member:
Jiawei Huang
Jingting Xu
Qiaochu Cong
Qiurong Ren

Contents

Problem Description

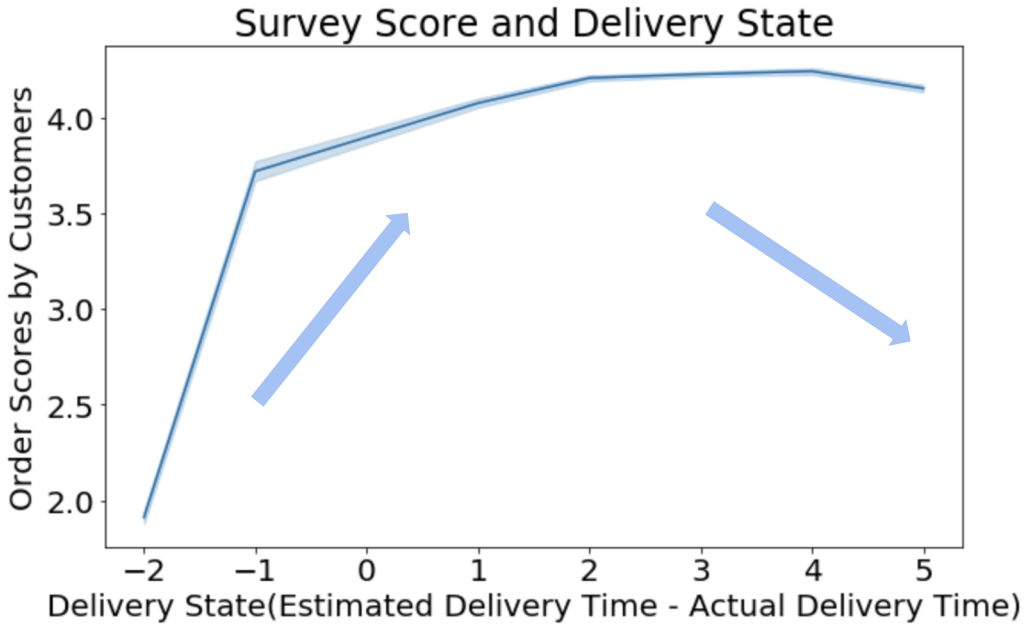
Explain why problem is interesting, what real-life application is being addressed

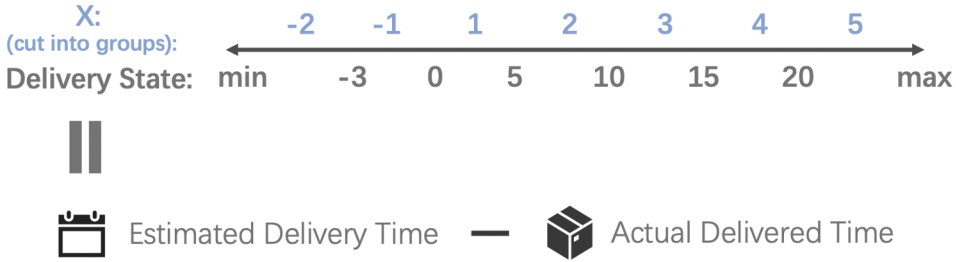




Problem Description

Explain why problem is interesting, what real-life application is being addressed





Problem Description

Explain why problem is interesting, what real-life application is being addressed

Problem:

Increasing delivery days → Lower scores
Less delays → Higher scores→ Then goes lower

Methods:

Existed information → Suitable estimated delivery days

Expected Outcome:

+ Higher scores
+ Improving customer experiences

Why Interesting:

Inaccurate estimate delivery time → Improve it

Analysis Task

type of task & how does task related to business problem

Features + **Labels**

Predict

New Est. Delivery Days

v.s.

Old Est. Delivery Days

Data Description - 1

data quality issues and how we fixed it

1. In ‘geolocation_dataset’, some of city names are both in English and Portuguese.

```
geolocation_dataset['geo_city'].value_counts()

sao paulo                135800
rio de janeiro            62151
belo horizonte            27805
são paulo                 24918
curitiba                  16593
                          ...
realeza (manhuacu)            1
contendas do sincorá          1
taquaral de goiás             1
jacare (cabreuva)             1
muribeca                      1
Name: geo_city, Length: 8011, dtype: int64
```

Purchased a Brazil zip codes dataset

2. In ‘products_dataset’, the categories are various and hard to be grouped

```
products_catagory_name['product_category_name'].unique()

array(['beleza_saude', 'informatica_acessorios', 'automotivo',
       'cama_mesa_banho', 'moveis_decoracao', 'esporte_lazer',
       'perfumaria', 'utilidades_domesticas', 'telefonia',
       'relogios_presentes', 'alimentos_bebidas', 'bebes', 'papelaria',
       'tablets_impressao_imagem', 'brinquedos', 'telefonia_fixa',
       'ferramentas_jardim', 'fashion_bolsas_e_acessorios',
       'eletroportateis', 'consoles_games', 'audio', 'fashion_calcados',
       'cool_stuff', 'malas_acessorios', 'climatizacao',
       'construcao_ferramentas_construcao',
       'moveis_cozinha_area_de_servico_jantar_e_jardim',
       'construcao_ferramentas_jardim', 'fashion_roupa_masculina',
       'pet_shop', 'moveis_escritorio', 'market_place', 'eletronicos',
       'eletrodomesticos', 'artigos_de_festas', 'casa_conforto',
       'construcao_ferramentas_ferramentas', 'agro_industria_e_comercio',
       'moveis_colchao_e_estofado', 'livros_tecnicos', 'casa_construcao',
       'instrumentos_musicais', 'moveis_sala',
       'construcao_ferramentas_iluminacao',
       'industria_comercio_e_negocios', 'alimentos', 'artes',
       'moveis_quarto', 'livros_interesse_geral',
       'construcao_ferramentas_seguranca',
       'fashion_underwear_e_moda_praia', 'fashion_esporte',
       'sinalizacao_e_seguranca', 'pcs', 'artigos_de_natal',
       'fashion_roupa_feminina', 'eletrodomesticos_2',
       'livros_importados', 'bebidas', 'cine_foto', 'la_cuisine',
       'musica', 'casa_conforto_2', 'portateis_casa_forno_e_cafe',
       'cds_dvds_musicais', 'dvds_blu_ray', 'flores',
       'artes_e_artesanato', 'fraldas_higiene',
       'fashion_roupa_infanto_juvenil', 'seguros_e_servicos'],
      dtype=object)
```

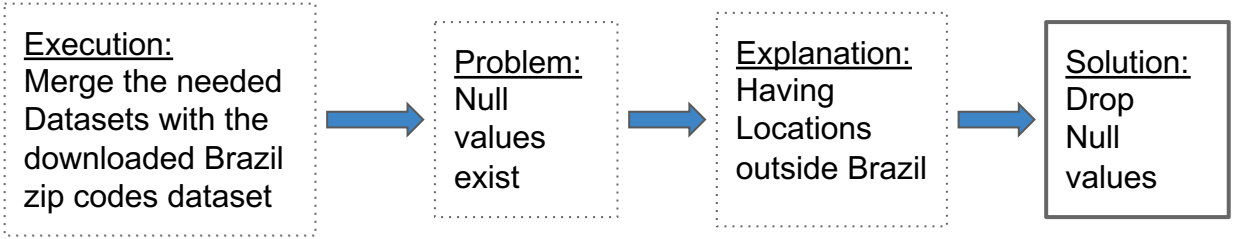
```
final['Category'].value_counts()

Furniture                   23077
Electric_appliance          19762
Home_and_office             19627
Sports_and_entertainment    14365
Beauty                      13311
Fashion                     10000
Baby                         7337
Auto                         4293
Food_drinks                  1508
Commerce                      514
Name: Category, dtype: int64
```

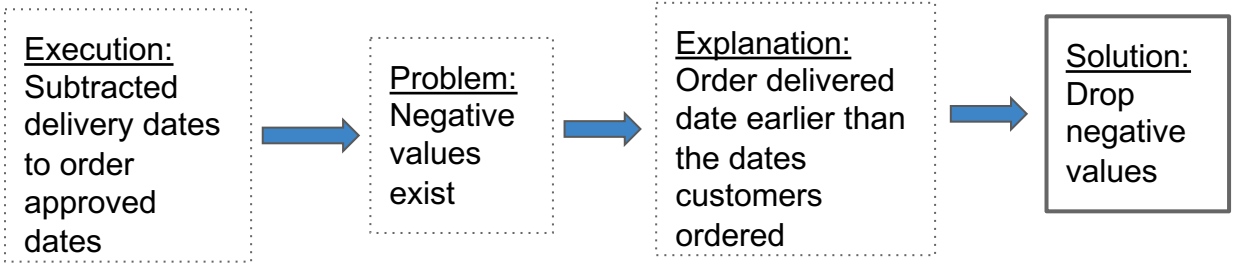
Data Description - 1

data quality issues and how we fixed it

3. Some locations are outliers which beyond the area of Brazil.



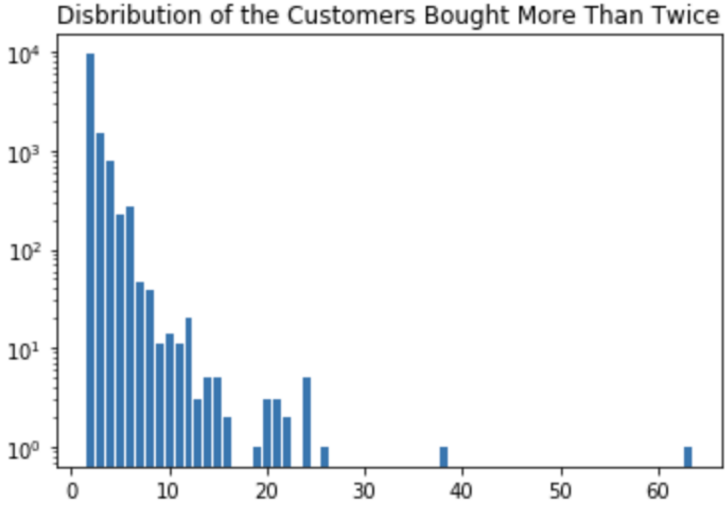
4. Some delivery dates are earlier than the time when the orders were approved



Data Description - 2

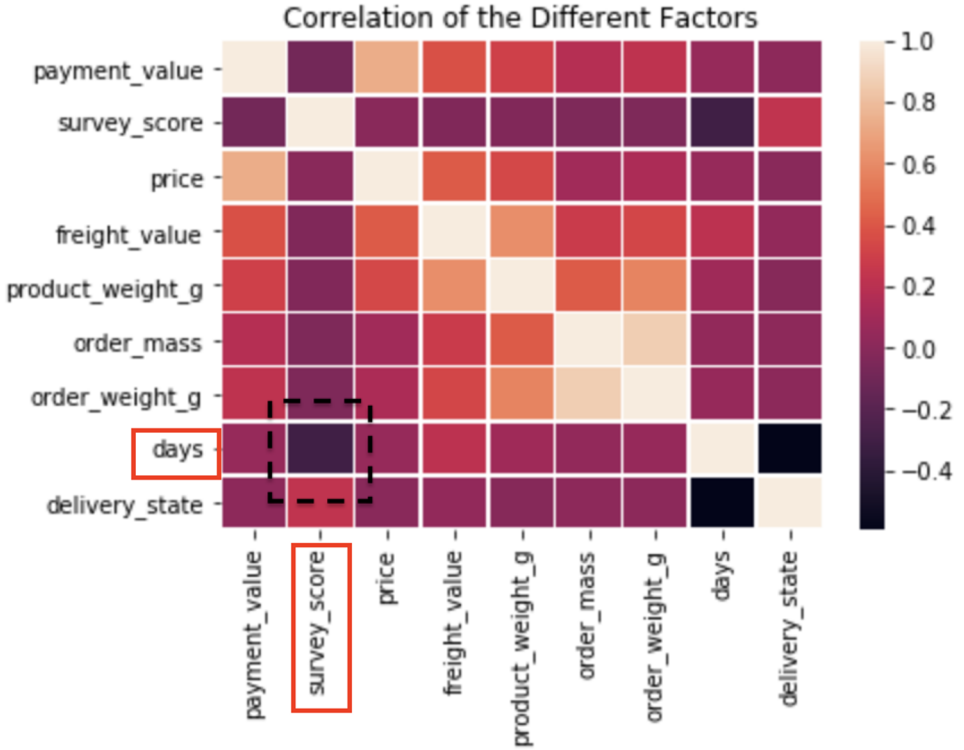
characteristics of the dataset, plots

	Purchased more than once	Purchased once
Number	31322	82495



Data Description - 2

characteristics of the dataset, plots



Data Description - 2

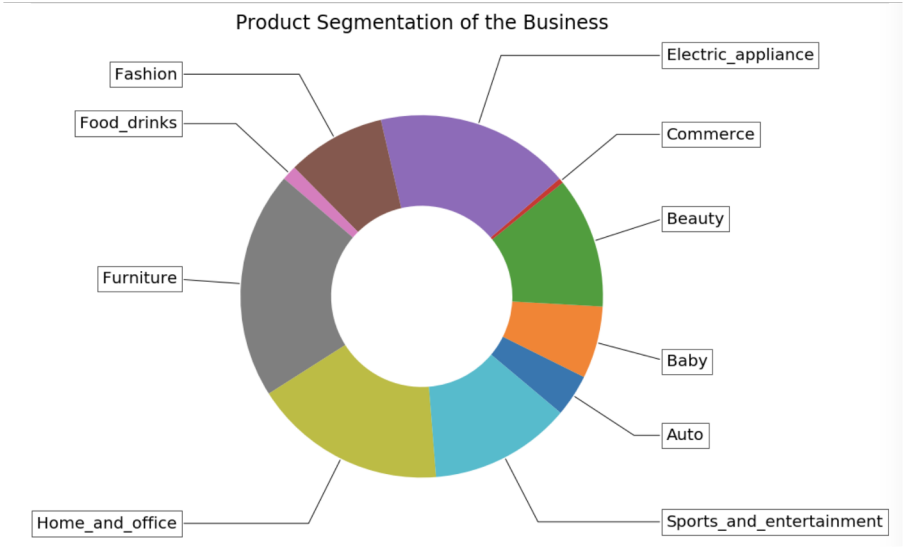
characteristics of the dataset, plots

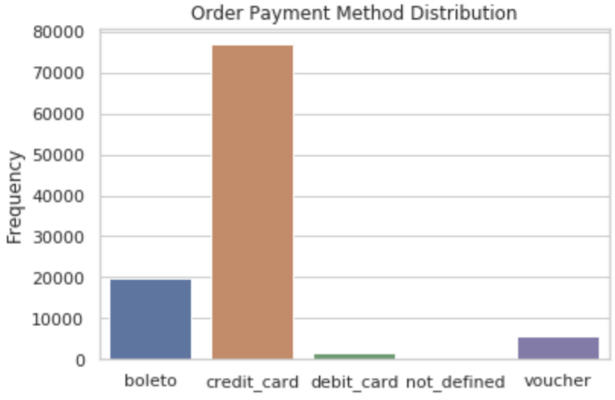
Word Cloud of Customer Review Contents

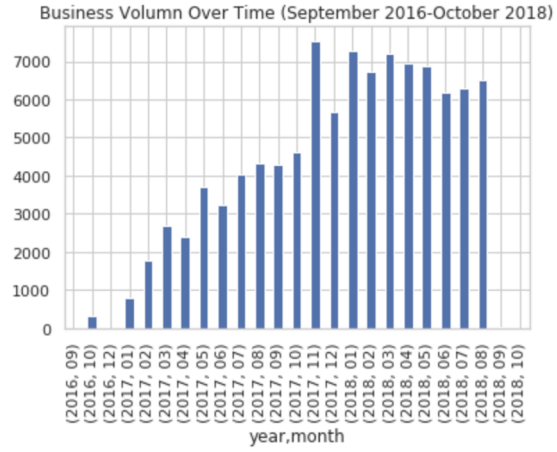
Word Cloud of Customer Review Title

Data Description - 2

characteristics of the dataset, plots

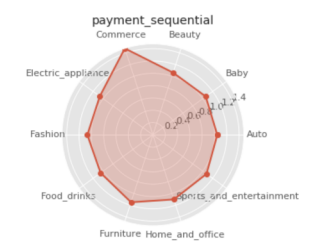


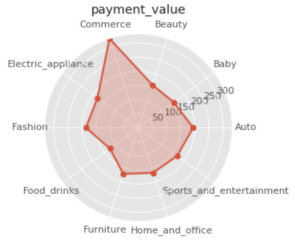


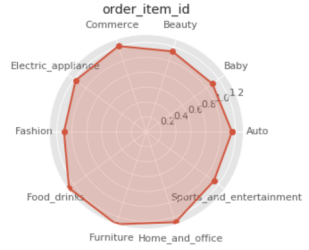


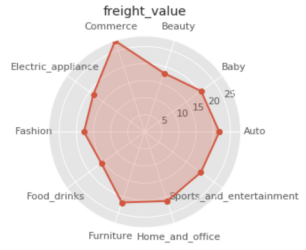
Data Description - 2

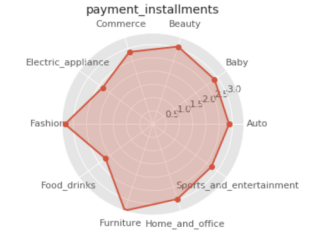
characteristics of the dataset, plots

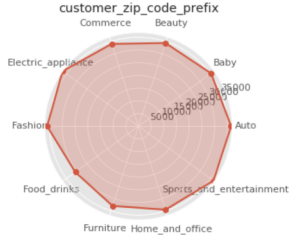


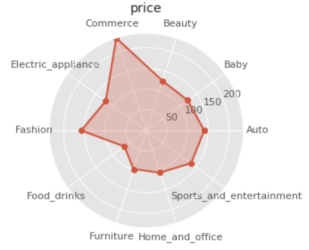


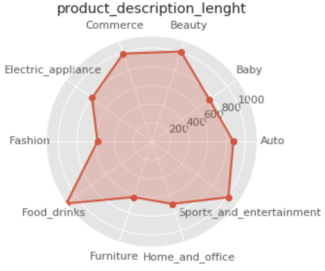


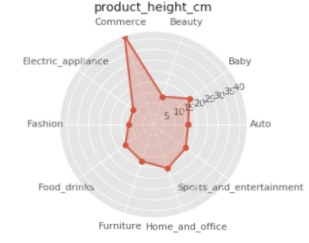


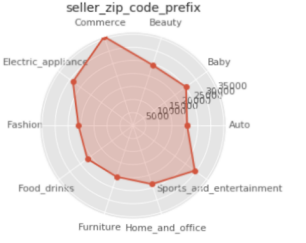


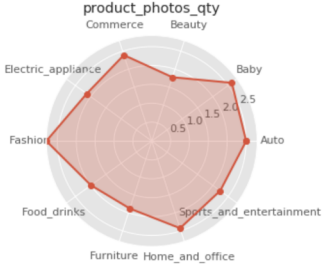


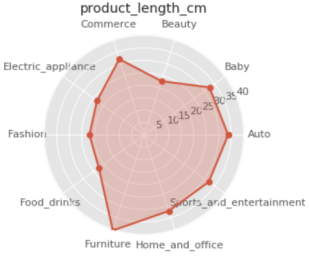


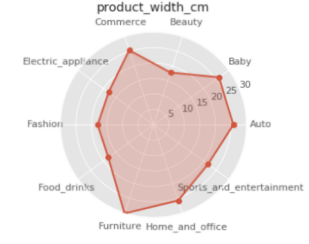


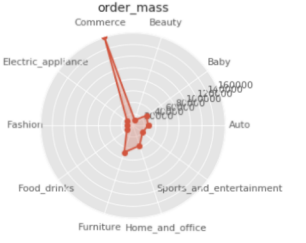






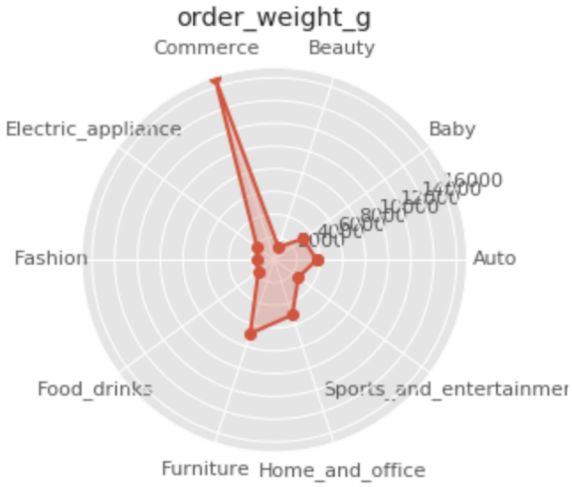


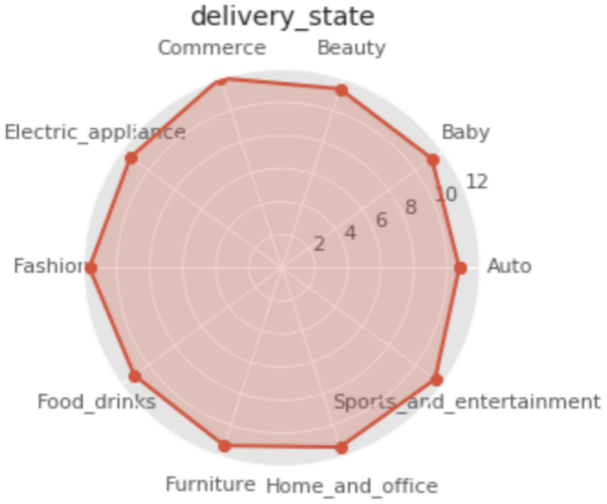


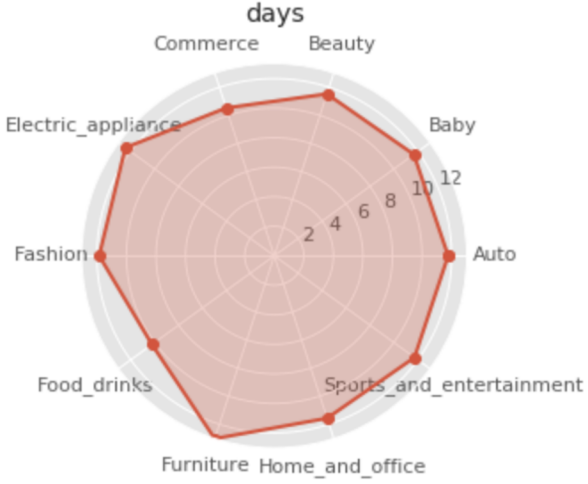


Data Description - 2

characteristics of the dataset, plots

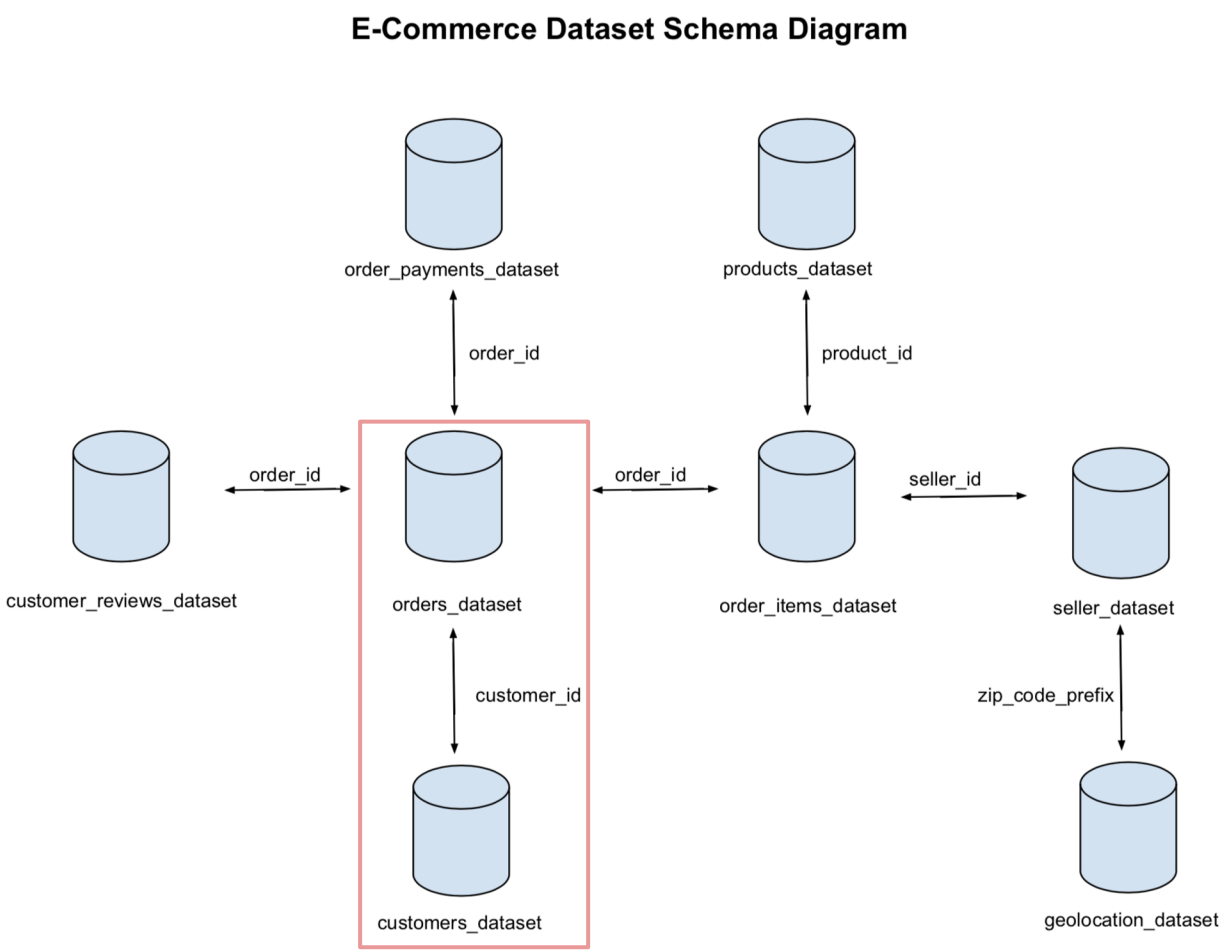






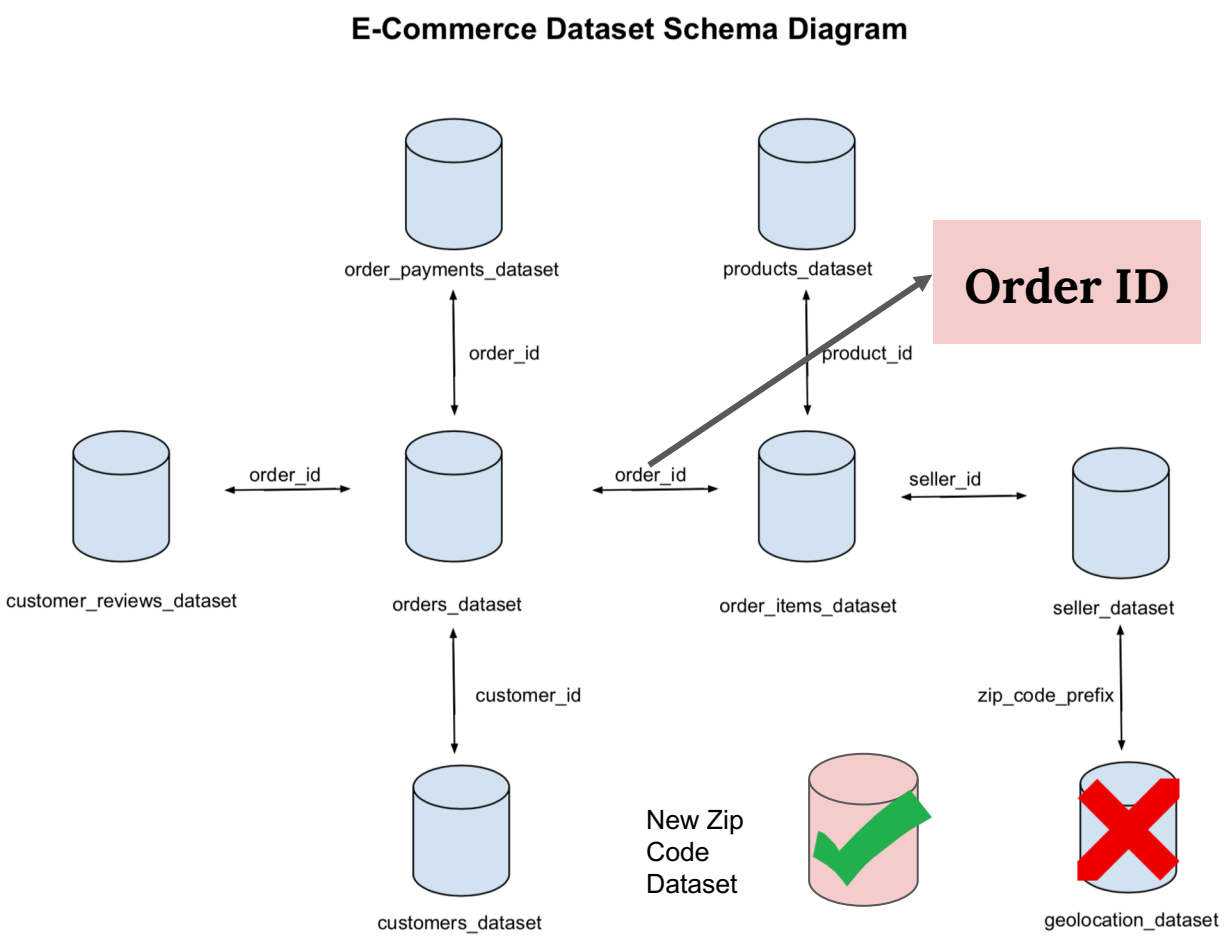
Data Preparation

data cleaning steps, features used



Data Preparation

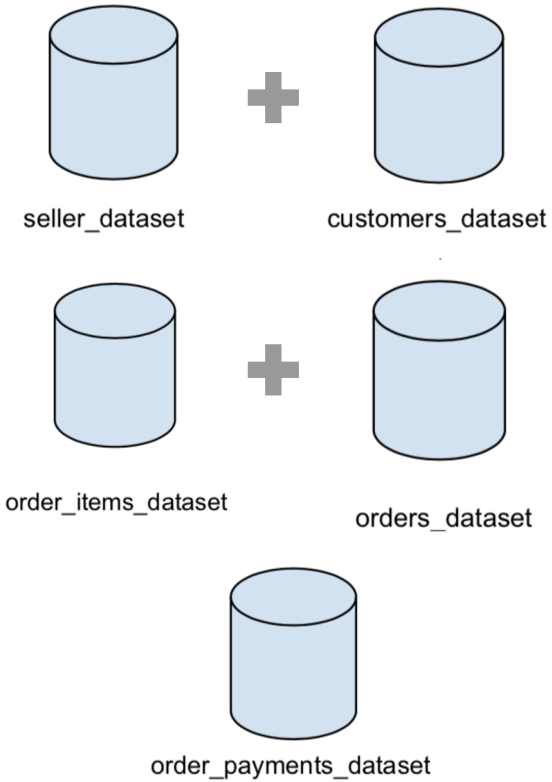
data cleaning steps, features used



Data Preparation

data cleaning steps, features used

Data Used:



New Features:

Same Region ?

Euclidean Distance

Order Volume

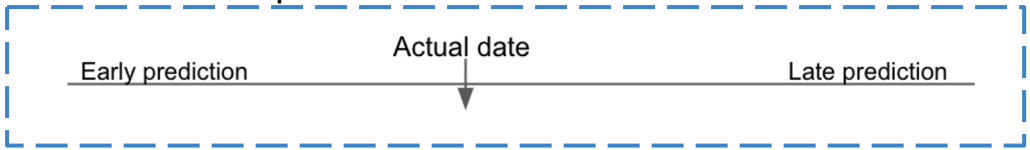
Order Weight

Actual D Days

Expected D Days

Planned Analysis

- Training & Testing Dataset: 80%,20%
- Regression: predict a more accurate and better delivery time,and provide a delivery time which suits for customers expectations, and help increase the customer's experience.
  a. Linear regression
  b. Extra Trees regression
  c. Random Forest regression
  d. …
- Evaluation: Instead of evaluate if our prediction is the exact deliver days, we are going to compare our prediction to the original expected delivery days to see which one is better using following criteria:
  a. Which prediction is closer to real delivery days
  b. Later prediction is better than earlier



- Packages for analysis: numpy, pandas, seaborn, matplotlib, sklearn, ExtraTreesRegressor, Randomforest