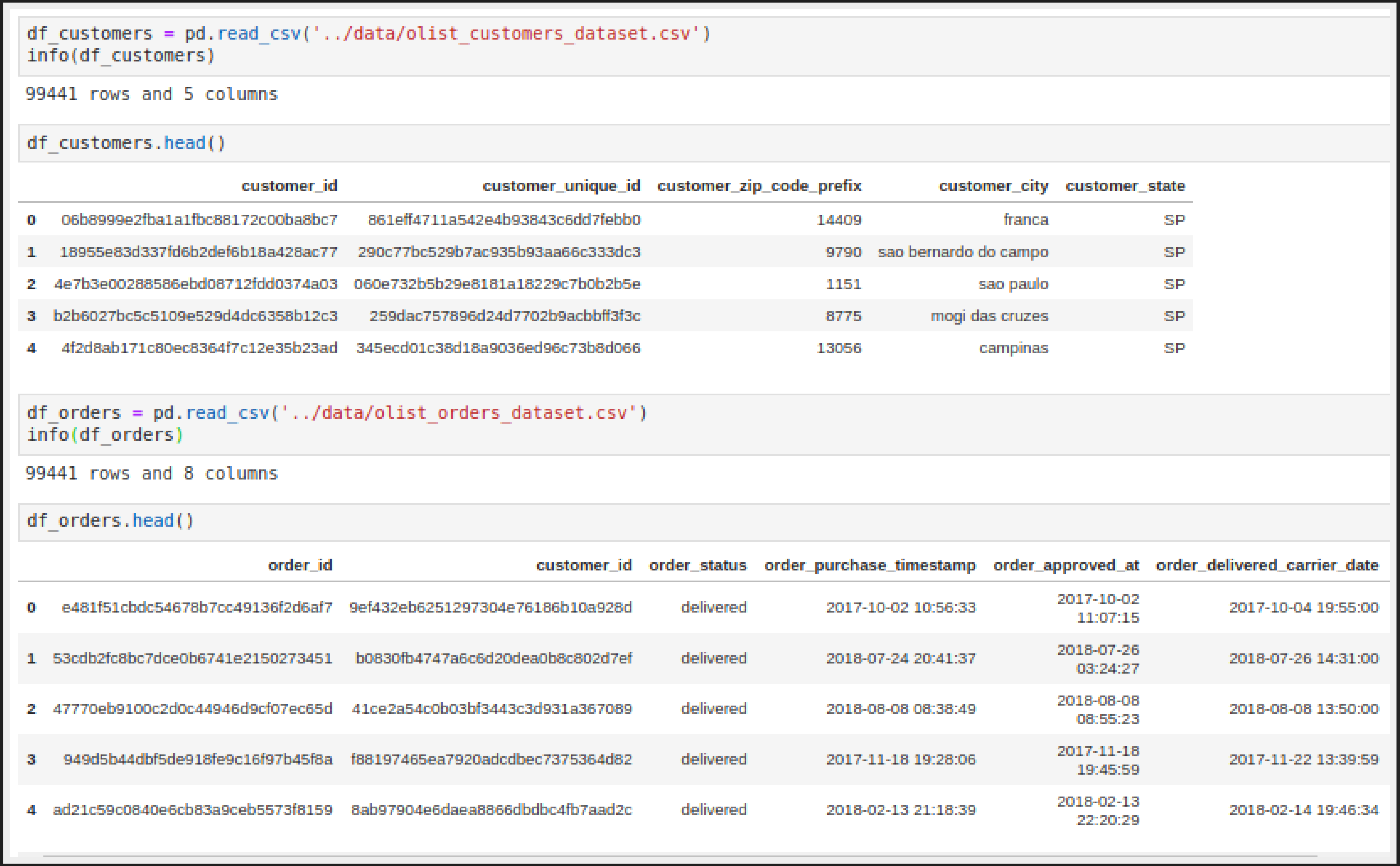
# SISTEMI BAZA PODATAKA

Aleksandar Škrbić

# PODACI

- Skup podataka je preuzet sa sajta Kaggle
- Olist je brazilska kompanija koja se bavi online prodajom
- Skup podataka sadrži infromacije o ~100k porudžbina u periodu od 2016-2018, ali najviše je primeraka iz 2018.
- Kompletan skup podataka se sastoji od 9 csv fajlova

- Struktura customers i orders fajlova koji sadrže podatke o korisnicima i porudžbinama
- Polje customer_id iz fajla custmers je referenca ka customer_id polju u orders fajlu



```
df_customers = pd.read_csv('../data/olist_customers_dataset.csv')
info(df_customers)
```

```
99441 rows and 5 columns
```

```
df_customers.head()
```

|  | customer_id | customer_unique_id | customer_zip_code_prefix | customer_city | customer_state |
|---|---|---|---|---|---|
| **0** | 06b8999e2fba1a1fbc88172c00ba8bc7 | 861eff4711a542e4b93843c6dd7febb0 | 14409 | franca | SP |
| **1** | 18955e83d337fd6b2def6b18a428ac77 | 290c77bc529b7ac935b93aa66c333dc3 | 9790 | sao bernardo do campo | SP |
| **2** | 4e7b3e00288586ebd08712fdd0374a03 | 060e732b5b29e8181a18229c7b0b2b5e | 1151 | sao paulo | SP |
| **3** | b2b6027bc5c5109e529d4dc6358b12c3 | 259dac757896d24d7702b9acbbff3f3c | 8775 | mogi das cruzes | SP |
| **4** | 4f2d8ab171c80ec8364f7c12e35b23ad | 345ecd01c38d18a9036ed96c73b8d066 | 13056 | campinas | SP |

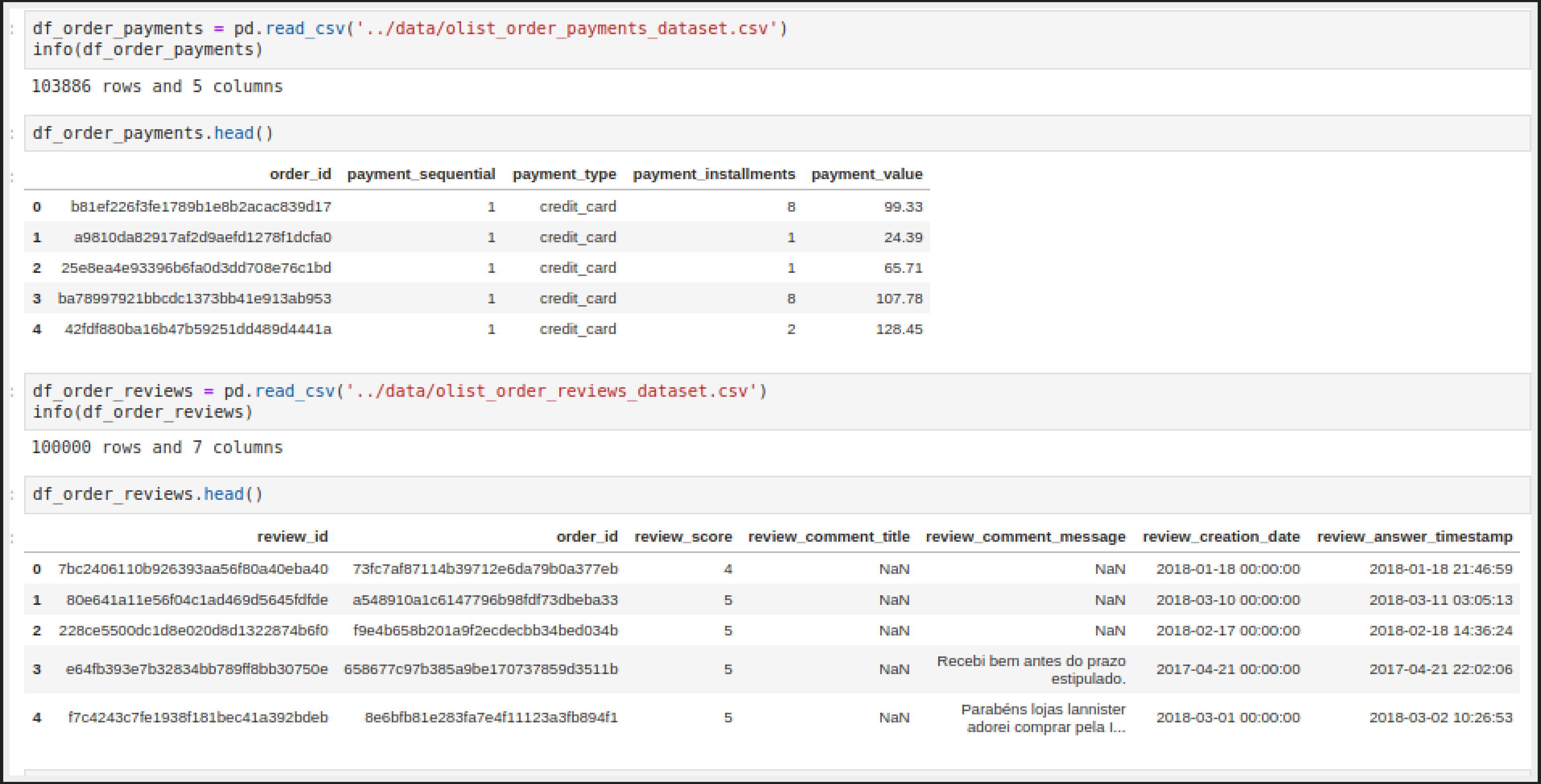
```
df_orders = pd.read_csv('../data/olist_orders_dataset.csv')
info(df_orders)
```

```
99441 rows and 8 columns
```

```
df_orders.head()
```

|  | order_id | customer_id | order_status | order_purchase_timestamp | order_approved_at | order_delivered_carrier_date |
|---|---|---|---|---|---|---|
| **0** | e481f51cbdc54678b7cc49136f2d6af7 | 9ef432eb6251297304e76186b10a928d | delivered | 2017-10-02 10:56:33 | 2017-10-02 11:07:15 | 2017-10-04 19:55:00 |
| **1** | 53cdb2fc8bc7dce0b6741e2150273451 | b0830fb4747a6c6d20dea0b8c802d7ef | delivered | 2018-07-24 20:41:37 | 2018-07-26 03:24:27 | 2018-07-26 14:31:00 |
| **2** | 47770eb9100c2d0c44946d9cf07ec65d | 41ce2a54c0b03bf3443c3d931a367089 | delivered | 2018-08-08 08:38:49 | 2018-08-08 08:55:23 | 2018-08-08 13:50:00 |
| **3** | 949d5b44dbf5de918fe9c16f97b45f8a | f88197465ea7920adcdbec7375364d82 | delivered | 2017-11-18 19:28:06 | 2017-11-18 19:45:59 | 2017-11-22 13:39:59 |
| **4** | ad21c59c0840e6cb83a9ceb5573f8159 | 8ab97904e6daea8866dbdbc4fb7aad2c | delivered | 2018-02-13 21:18:39 | 2018-02-13 22:20:29 | 2018-02-14 19:46:34 |

- Struktura fajlova koji sadrže dodatne detalje o porudžbinama, kao sto su način plaćanja i recenzija
- Fajlovi se mogu spojiti putem order_id polja



```
df_order_payments = pd.read_csv('../data/olist_order_payments_dataset.csv')
info(df_order_payments)
```

```
103886 rows and 5 columns
```

```
df_order_payments.head()
```

| | order_id | payment_sequential | payment_type | payment_installments | payment_value |
|---|---|---|---|---|---|
| 0 | b81ef226f3fe1789b1e8b2acac839d17 | 1 | credit_card | 8 | 99.33 |
| 1 | a9810da82917af2d9aefd1278f1dcfa0 | 1 | credit_card | 1 | 24.39 |
| 2 | 25e8ea4e93396b6fa0d3dd708e76c1bd | 1 | credit_card | 1 | 65.71 |
| 3 | ba78997921bbcdc1373bb41e913ab953 | 1 | credit_card | 8 | 107.78 |
| 4 | 42fdf880ba16b47b59251dd489d4441a | 1 | credit_card | 2 | 128.45 |

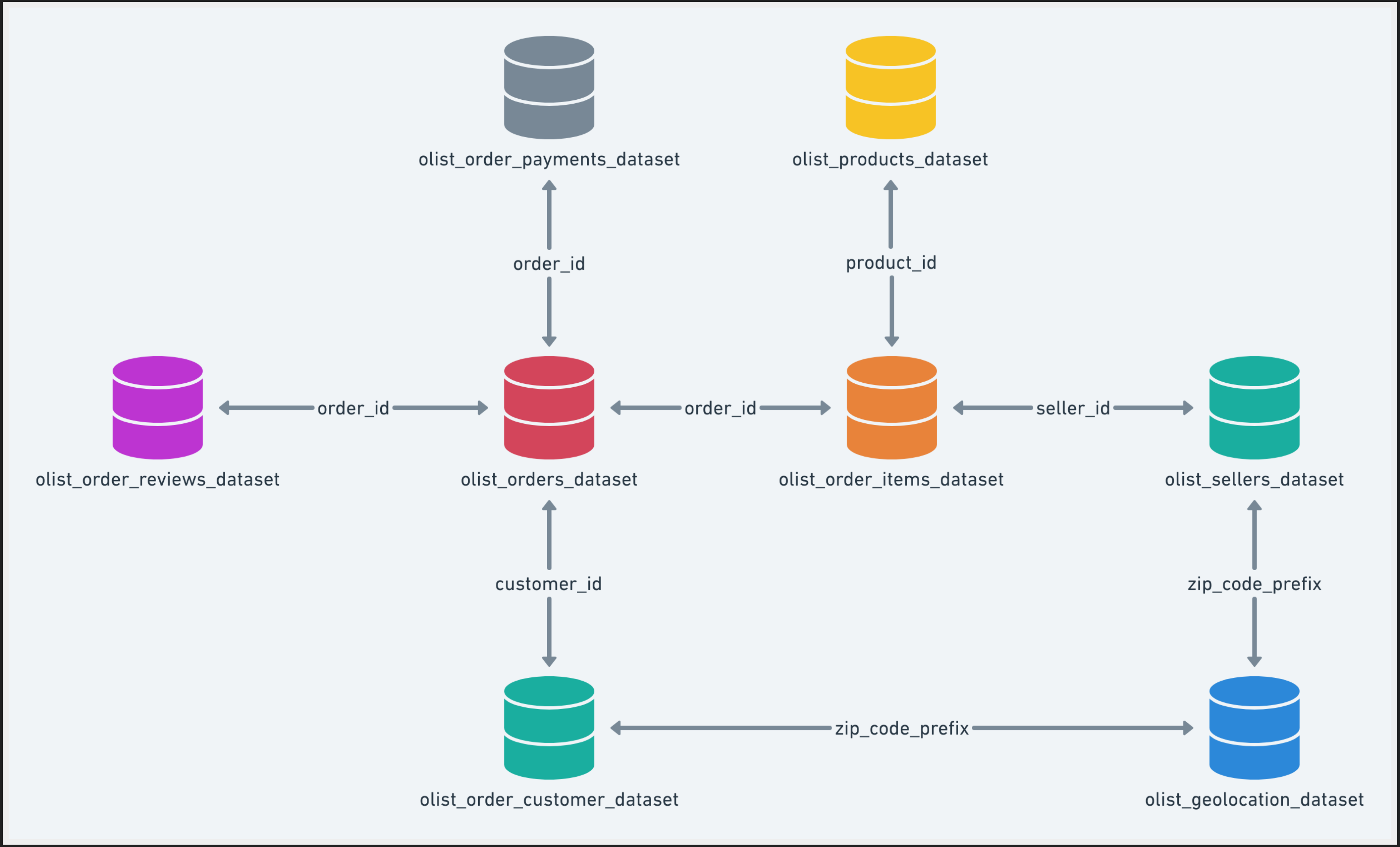
```
df_order_reviews = pd.read_csv('../data/olist_order_reviews_dataset.csv')
info(df_order_reviews)
```

```
100000 rows and 7 columns
```

```
df_order_reviews.head()
```

| | review_id | order_id | review_score | review_comment_title | review_comment_message | review_creation_date | review_answer_timestamp |
|---|---|---|---|---|---|---|---|
| 0 | 7bc2406110b926393aa56f80a40eba40 | 73fc7af87114b39712e6da79b0a377eb | 4 | NaN | NaN | 2018-01-18 00:00:00 | 2018-01-18 21:46:59 |
| 1 | 80e641a11e56f04c1ad469d5645fdfde | a548910a1c6147796b98fdf73dbeba33 | 5 | NaN | NaN | 2018-03-10 00:00:00 | 2018-03-11 03:05:13 |
| 2 | 228ce5500dc1d8e020d8d1322874b6f0 | f9e4b658b201a9f2ecdecbb34bed034b | 5 | NaN | NaN | 2018-02-17 00:00:00 | 2018-02-18 14:36:24 |
| 3 | e64fb393e7b32834bb789ff8bb30750e | 658677c97b385a9be170737859d3511b | 5 | NaN | Recebi bem antes do prazo estipulado. | 2017-04-21 00:00:00 | 2017-04-21 22:02:06 |
| 4 | f7c4243c7fe1938f181bec41a392bdeb | 8e6bfb81e283fa7e4f11123a3fb894f1 | 5 | NaN | Parabéns lojas lannister adorei comprar pela I... | 2018-03-01 00:00:00 | 2018-03-02 10:26:53 |

- Prikaz kompletne šeme fajlova, koja pokazuje na koji način je moguće spojiti fajlove

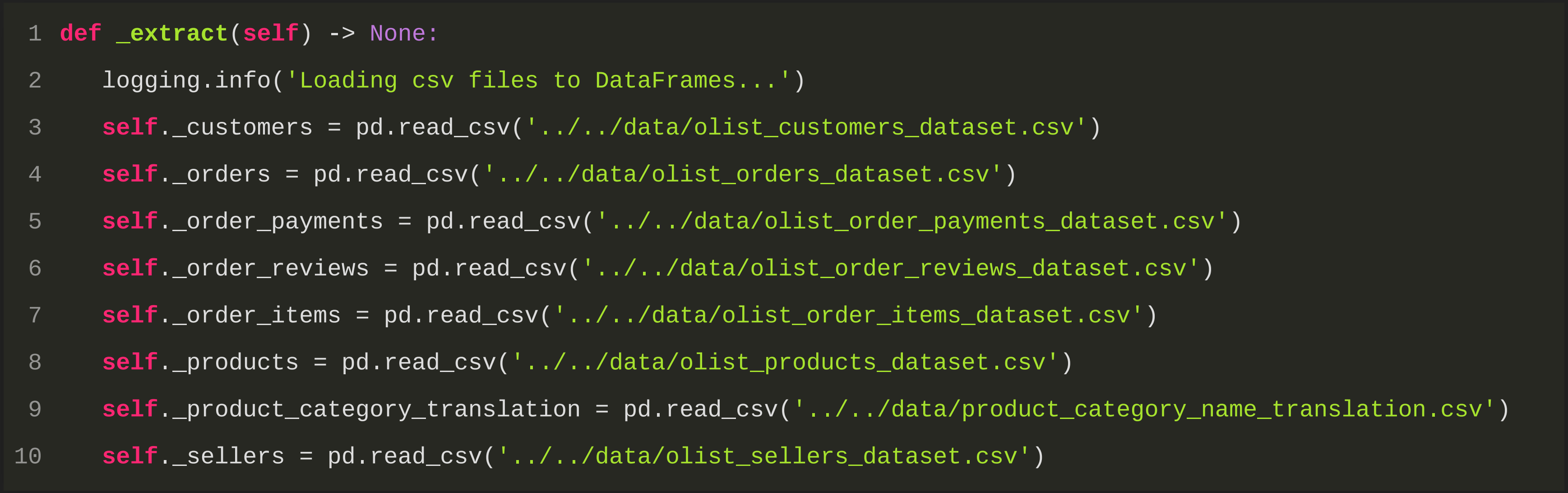


## OPIS PROCESA OBRADE PODATAKA

Korišćenje tehnologije:

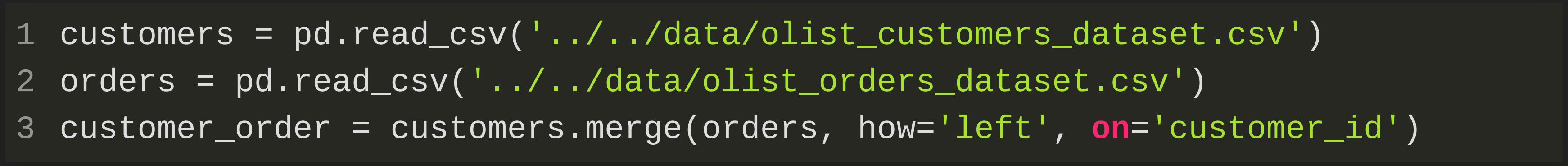
- Python
- Pandas - Python biblioteka za manipulaciju podacima(čišćenje, spajanje, agregacija...)
- PyMongo - Python klijent za MongoDB

- Primer učitavanja fajlova u DataFrame - struktura podataka koju nam obezbeđuje Pandas
- DataFrame je veoma sličan tabelama u relacionoj bazi podataka



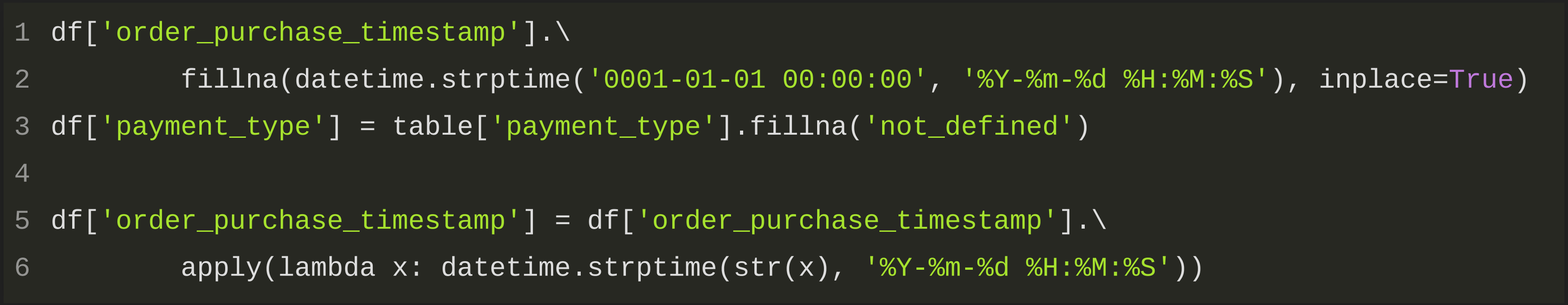
```python
def _extract(self) -> None:
   logging.info('Loading csv files to DataFrames...')
   self._customers = pd.read_csv('../../data/olist_customers_dataset.csv')
   self._orders = pd.read_csv('../../data/olist_orders_dataset.csv')
   self._order_payments = pd.read_csv('../../data/olist_order_payments_dataset.csv')
   self._order_reviews = pd.read_csv('../../data/olist_order_reviews_dataset.csv')
   self._order_items = pd.read_csv('../../data/olist_order_items_dataset.csv')
   self._products = pd.read_csv('../../data/olist_products_dataset.csv')
   self._product_category_translation = pd.read_csv('../../data/product_category_name_translation.csv')
   self._sellers = pd.read_csv('../../data/olist_sellers_dataset.csv')
```

- **Primer učitavanja i spajanja dva DataFrame-a**



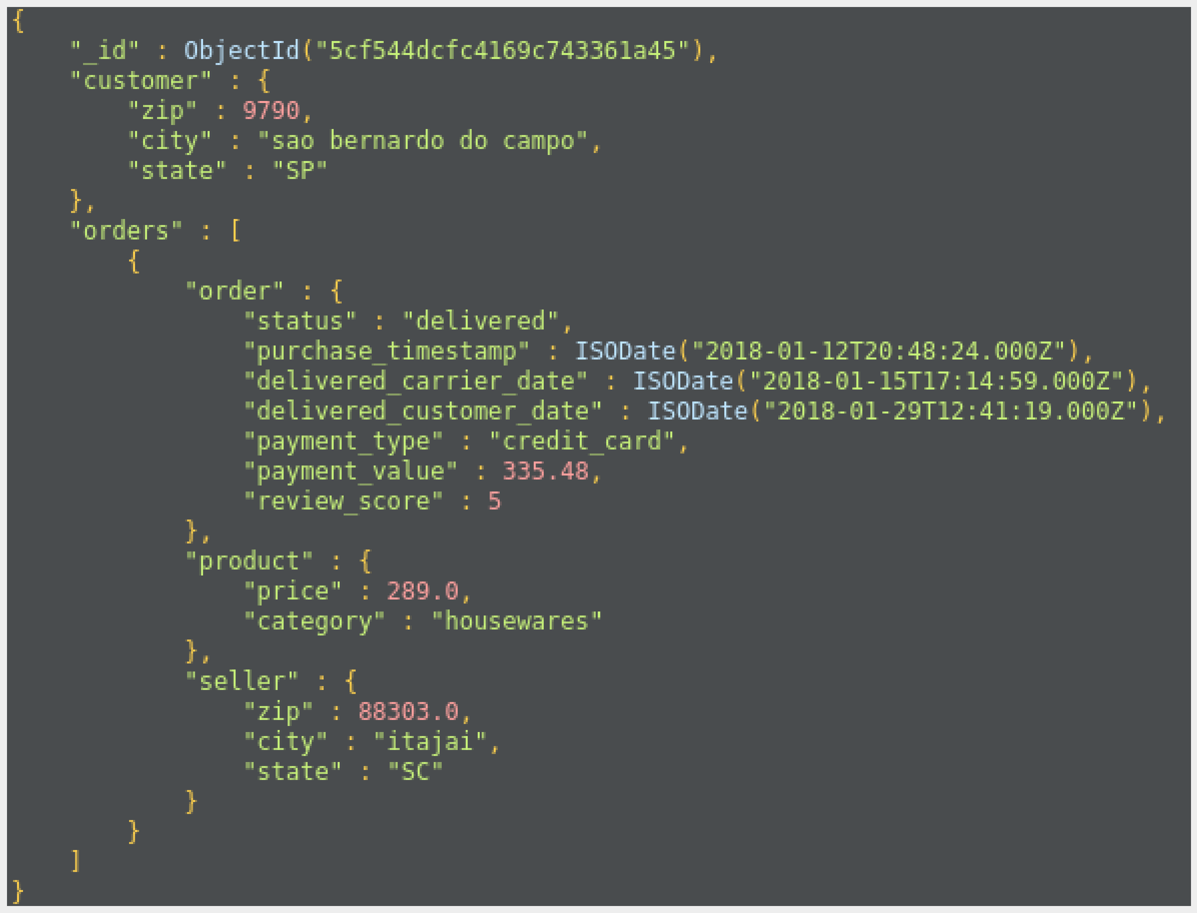
```
customers = pd.read_csv('../../data/olist_customers_dataset.csv')
orders = pd.read_csv('../../data/olist_orders_dataset.csv')
customer_order = customers.merge(orders, how='left', on='customer_id')
```

- **Kako rešiti problem nedostajućih vrednosti u tabeli?**



```
df['order_purchase_timestamp'].\
        fillna(datetime.strptime('0001-01-01 00:00:00', '%Y-%m-%d %H:%M:%S'), inplace=True)
df['payment_type'] = table['payment_type'].fillna('not_defined')

df['order_purchase_timestamp'] = df['order_purchase_timestamp'].\
        apply(lambda x: datetime.strptime(str(x), '%Y-%m-%d %H:%M:%S'))
```

# FINALNA MONGO ŠEMA DOKUMENTA



```
{
    "_id" : ObjectId("5cf544dcfc4169c743361a45"),
    "customer" : {
        "zip" : 9790,
        "city" : "sao bernardo do campo",
        "state" : "SP"
    },
    "orders" : [
        {
            "order" : {
                "status" : "delivered",
                "purchase_timestamp" : ISODate("2018-01-12T20:48:24.000Z"),
                "delivered_carrier_date" : ISODate("2018-01-15T17:14:59.000Z"),
                "delivered_customer_date" : ISODate("2018-01-29T12:41:19.000Z"),
                "payment_type" : "credit_card",
                "payment_value" : 335.48,
                "review_score" : 5
            },
            "product" : {
                "price" : 289.0,
                "category" : "housewares"
            },
            "seller" : {
                "zip" : 88303.0,
                "city" : "itajai",
                "state" : "SC"
            }
        }
    ]
}
```

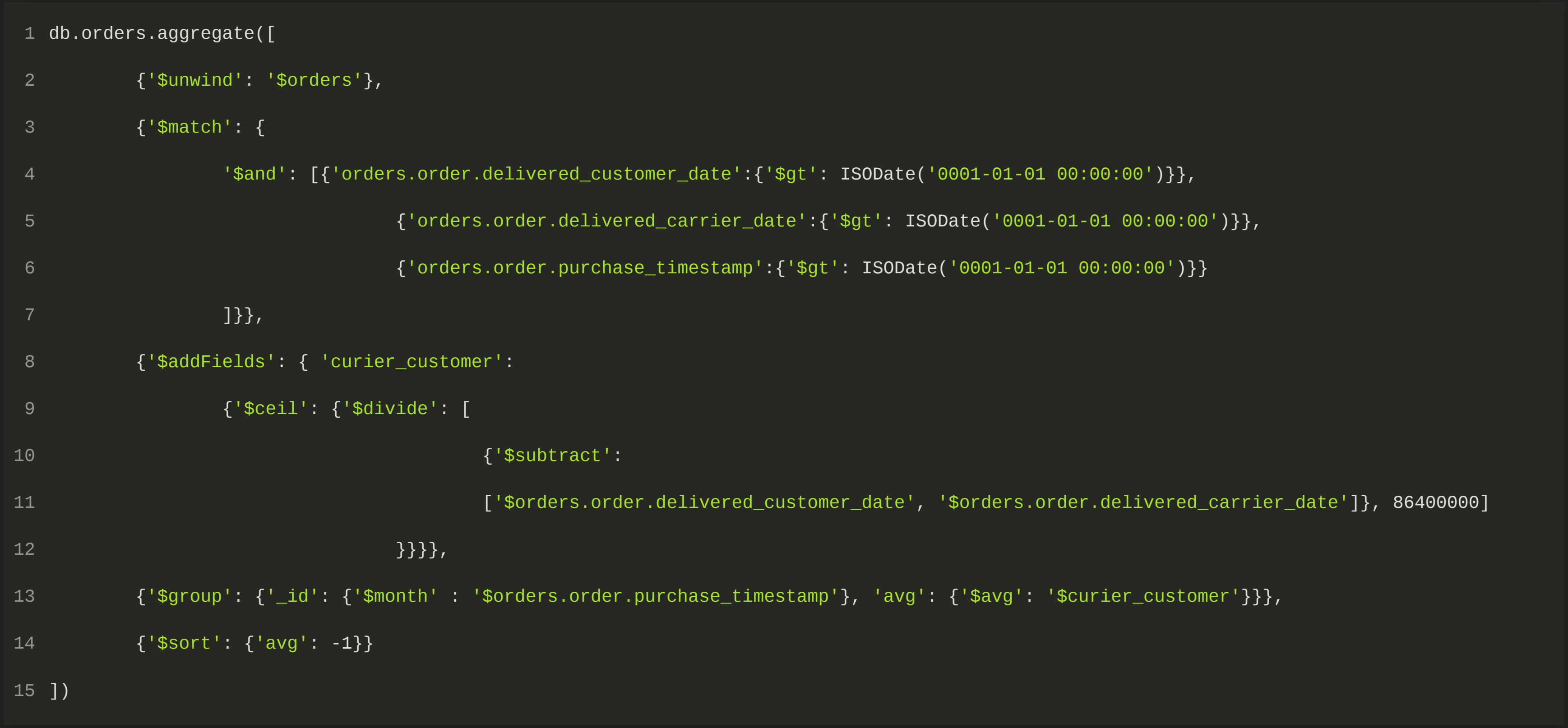
# PITANJA

- Broj porudžbina po gradu i državi
- Minimalna i maksimalna cena porudžbine
- Broj porudžbina po statusu porudžbine
- Koliko porudžbina je dobilo koju ocenu
- Prosečan broj porudžbina i prosečna potrošnja po mesecima
- Prosečna potrošnja po gradovima i državama
- Najpopularnije metode plaćanja
- Prosečna cena porudžbine po metodi plaćanja
- Top 10 najpopularnijih kategorija proizvoda
- Prosečno vreme čekanja od naručivanja do predaje kuriru
- Prosečno vreme čekanja od kurira do kupca
- Prosečno vreme čekanja od naručivanja do kupca

- **Prosečna potrošnja po metodu plaćanja**

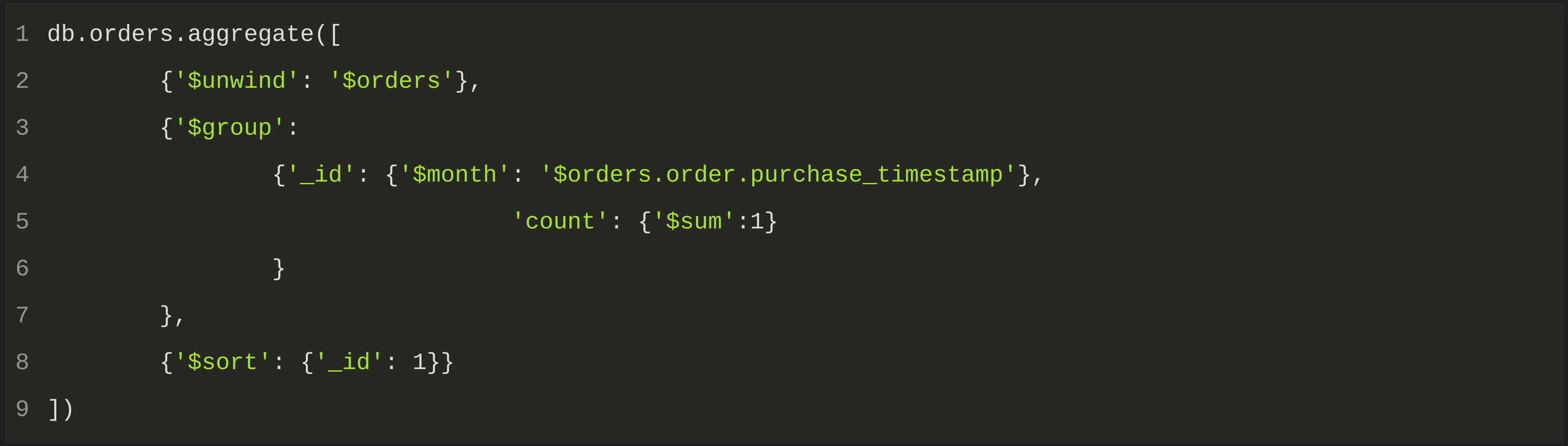
```
db.orders.aggregate(
        {'$unwind': '$orders'},
        {'$group':
                {'_id': '$orders.order.payment_type', 'avg': {'$avg': '$orders.order.payment_value'}}
        },
        {'$sort': {'avg': -1}},
)
```

- Prosečno vreme u danima koje treba kuriru da dostavi porudžbinu



```
db.orders.aggregate([
        {'$unwind': '$orders'},
        {'$match': {
                '$and': [{'orders.order.delivered_customer_date':{'$gt': ISODate('0001-01-01 00:00:00')}},
                                {'orders.order.delivered_carrier_date':{'$gt': ISODate('0001-01-01 00:00:00')}},
                                {'orders.order.purchase_timestamp':{'$gt': ISODate('0001-01-01 00:00:00')}}
                ]}},
        {'$addFields': { 'curier_customer':
                {'$ceil': {'$divide': [
                                        {'$subtract':
                                        ['$orders.order.delivered_customer_date', '$orders.order.delivered_carrier_date']}, 86400000]
                                }}}},
        {'$group': {'_id': {'$month' : '$orders.order.purchase_timestamp'}, 'avg': {'$avg': '$curier_customer'}}},
        {'$sort': {'avg': -1}}
])
```

- **Broj porudžbina po mesecima**



```
db.orders.aggregate([
        {'$unwind': '$orders'},
        {'$group':
                {'_id': {'$month': '$orders.order.purchase_timestamp'},
                                 'count': {'$sum':1}
                }
        },
        {'$sort': {'_id': 1}}
])
```

## REST API I VIZUALIZACIJA

U sklopu projekta, napisan je REST API preko kojeg se mogu dobiti odgovori u JSON formatu na prethodno definisana pitanja, kao i web aplikacija koja vizualizuje rezultate nekih upita

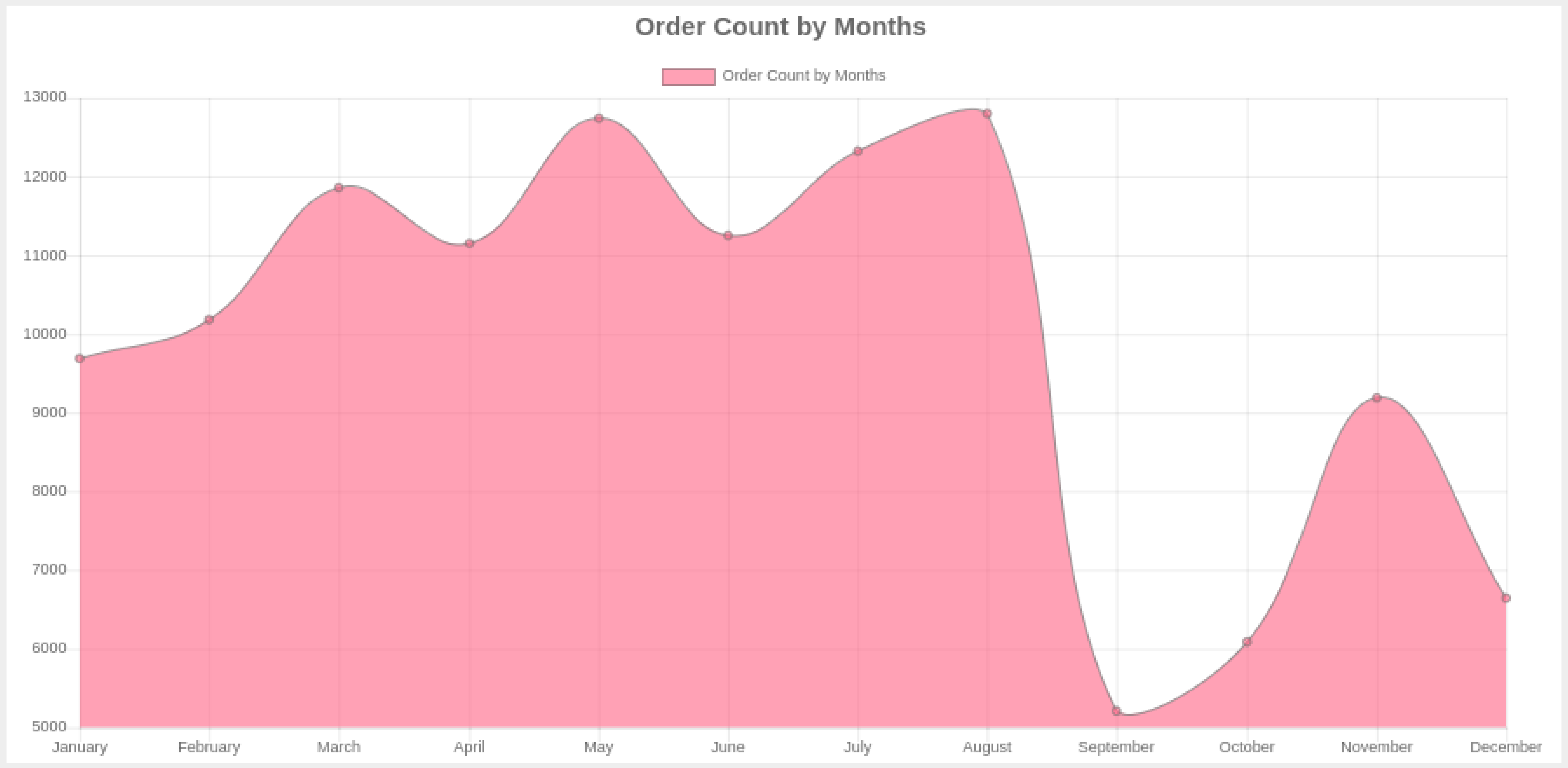
Korišćenje tehnologije:

- Python
- Javascript
- Flask - Web framework
- PyMongo - Python klijent za MongoDB
- Chart.js - Javascript biblioteka za vizualizaciju podataka

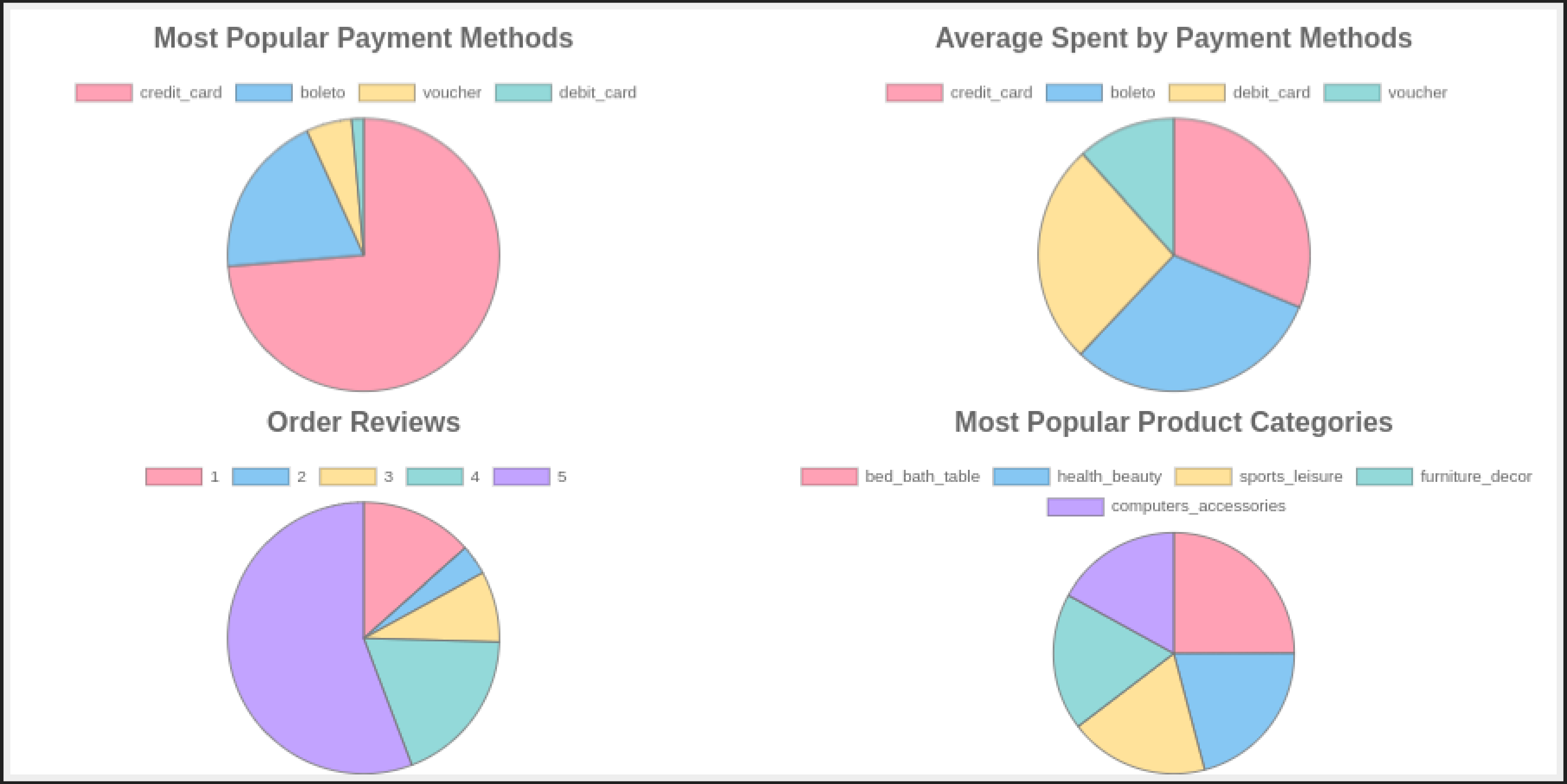
# PYMONGO PRIMER

```python
import pymongo

def get_most_popular_products() -> list:
  client = pymongo.MongoClient('mongodb://localhost:27017/') # konekcija na Mongo server
  db = client['sbp'] # selekcija baze
  collection = db['orders'] # selekcija kolekcije unutar baze

  pipline = [
        {'$unwind': '$orders'},
        {'$group': {'_id': '$orders.product.category', 'count': {'$sum': 1}}},
        {'$sort': {'count': -1}},
        {'$limit': 5}
  ]

  result = list(collection.aggregate(pipline))
```

# PRIMER VIZUALIZACIJE



# PRIMER VIZUALIZACIJE


Most Popular Payment Methods
credit_card
boleto
voucher
debit_card
Average Spent by Payment Methods
credit_card
boleto
debit_card
voucher
Order Reviews
Most Popular Product Categories
bed_bath_table
health_beauty
sports_leisure
furniture_decor
computers_accessories