PYTHON FOR DATASCIENCE

MAXENCE GUEGNOLLE - MARTIN DE COMBARIEU

BERLIN AIRBNB ANALYSIS

CAN WE PREDICT THE PRICE OF A NEW PLACE ?

SUMMARY

HOW WE PROCESSED

- Data
  - DataSet Selection
  - Data Exploration
  - Data Cleaning
- Features
  - Feature Analysis
  - Feature Engineering
- Model
- Conclusion

BERLIN AIRBNB DATA

kaggle

Price related

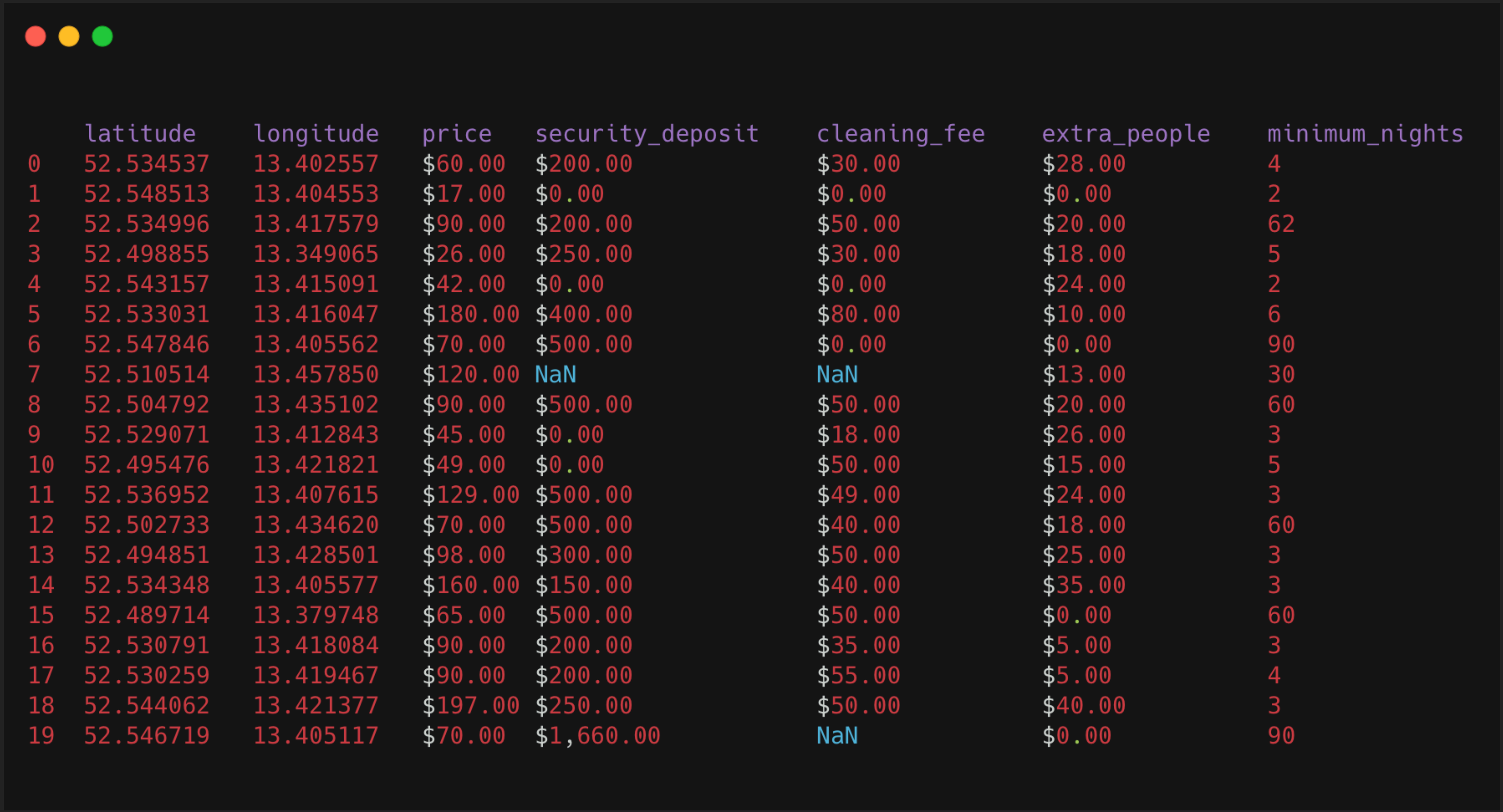
- Price
- Cleanning Fee
- Extra People
- Security Deposit

Amenities

Position

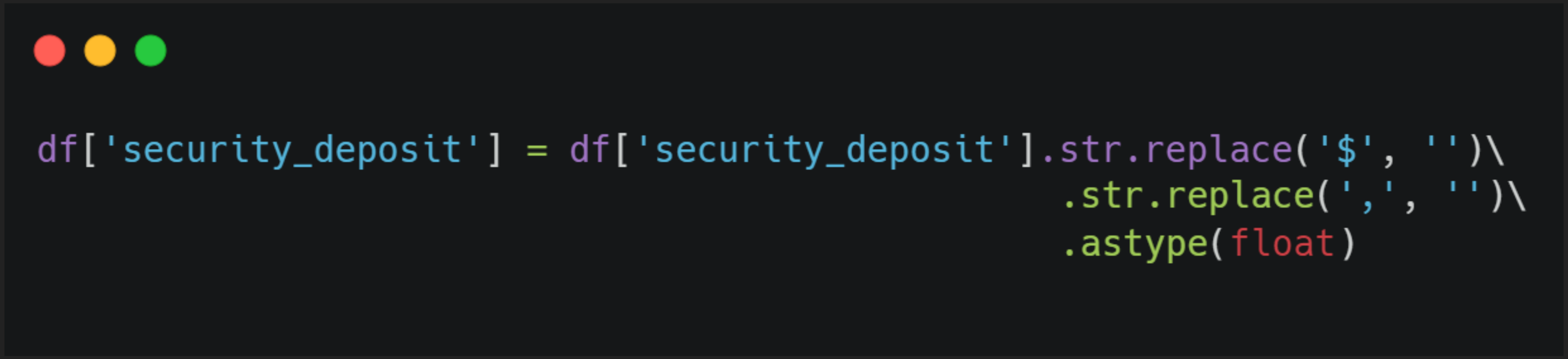
- Longitude
- Latitude

DATA EXPLORATION

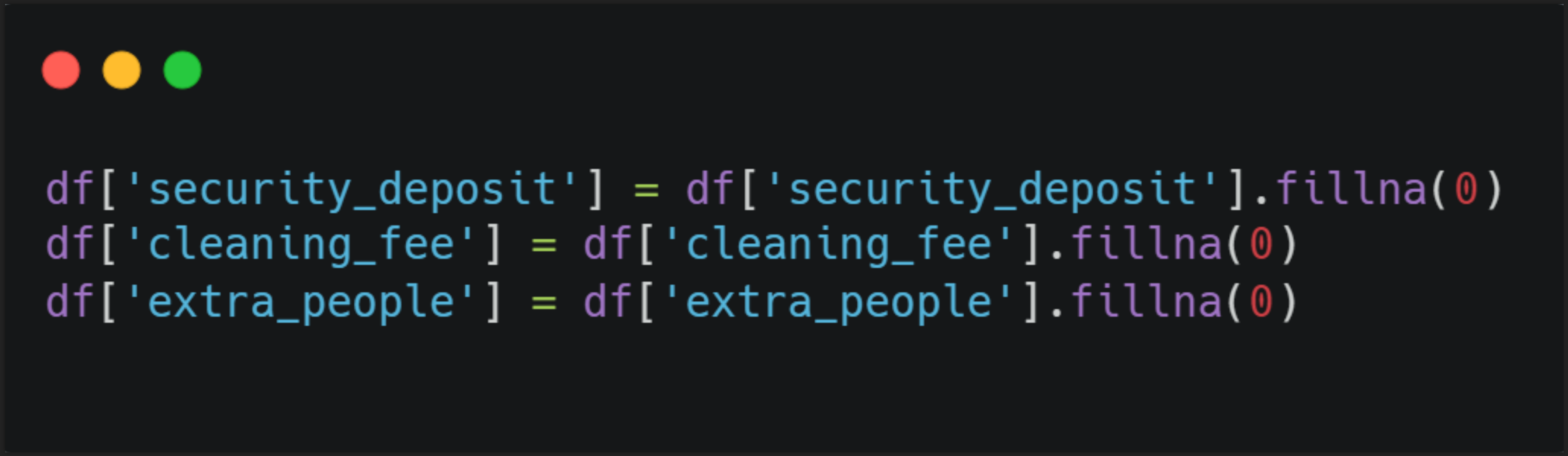


```
    latitude   longitude  price   security_deposit    cleaning_fee    extra_people    minimum_nights
0   52.534537  13.402557  $60.00  $200.00             $30.00          $28.00          4
1   52.548513  13.404553  $17.00  $0.00               $0.00           $0.00           2
2   52.534996  13.417579  $90.00  $200.00             $50.00          $20.00          62
3   52.498855  13.349065  $26.00  $250.00             $30.00          $18.00          5
4   52.543157  13.415091  $42.00  $0.00               $0.00           $24.00          2
5   52.533031  13.416047  $180.00 $400.00             $80.00          $10.00          6
6   52.547846  13.405562  $70.00  $500.00             $0.00           $0.00           90
7   52.510514  13.457850  $120.00 NaN                 NaN             $13.00          30
8   52.504792  13.435102  $90.00  $500.00             $50.00          $20.00          60
9   52.529071  13.412843  $45.00  $0.00               $18.00          $26.00          3
10  52.495476  13.421821  $49.00  $0.00               $50.00          $15.00          5
11  52.536952  13.407615  $129.00 $500.00             $49.00          $24.00          3
12  52.502733  13.434620  $70.00  $500.00             $40.00          $18.00          60
13  52.494851  13.428501  $98.00  $300.00             $50.00          $25.00          3
14  52.534348  13.405577  $160.00 $150.00             $40.00          $35.00          3
15  52.489714  13.379748  $65.00  $500.00             $50.00          $0.00           60
16  52.530791  13.418084  $90.00  $200.00             $35.00          $5.00           3
17  52.530259  13.419467  $90.00  $200.00             $55.00          $5.00           4
18  52.544062  13.421377  $197.00 $250.00             $50.00          $40.00          3
19  52.546719  13.405117  $70.00  $1,660.00           NaN             $0.00           90
```

DATA CLEANING – PRICE

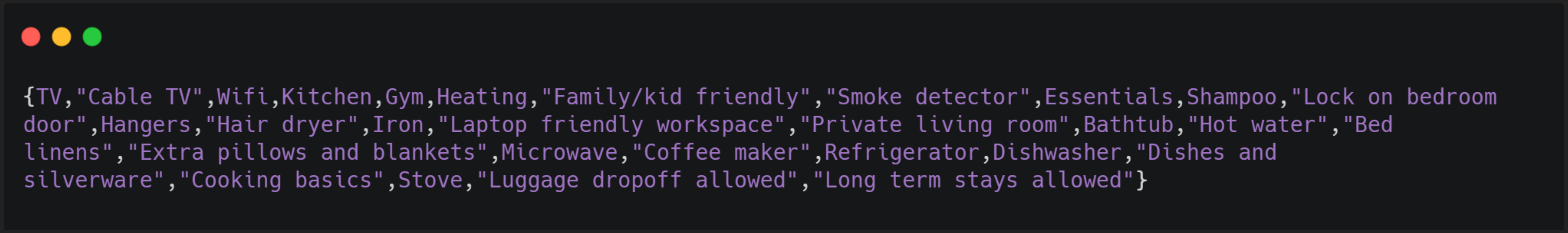


```
df['security_deposit'] = df['security_deposit'].str.replace('$', '')\
                                               .str.replace(',', '')\
                                               .astype(float)
```

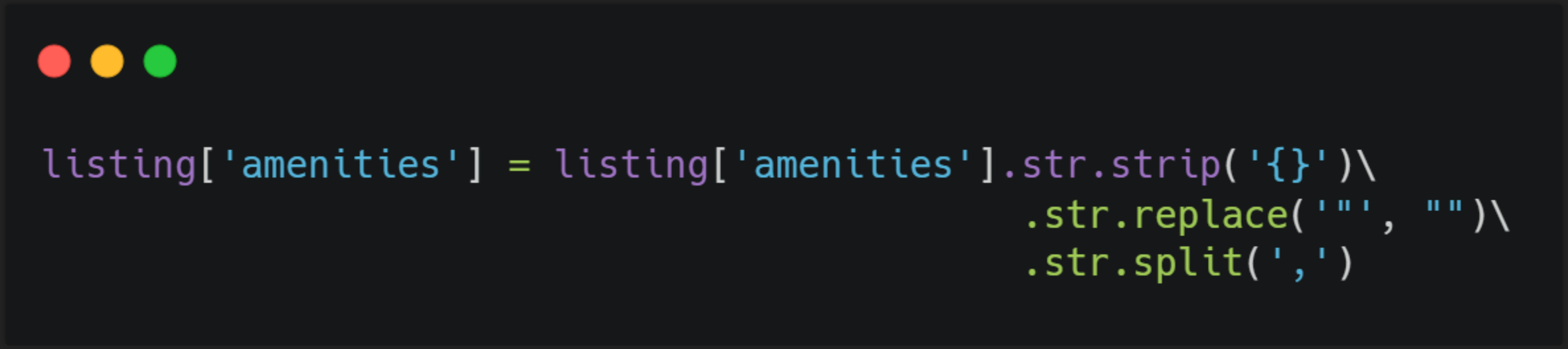


```
df['security_deposit'] = df['security_deposit'].fillna(0)
df['cleaning_fee'] = df['cleaning_fee'].fillna(0)
df['extra_people'] = df['extra_people'].fillna(0)
```

DATA CLEANING - AMENITIES

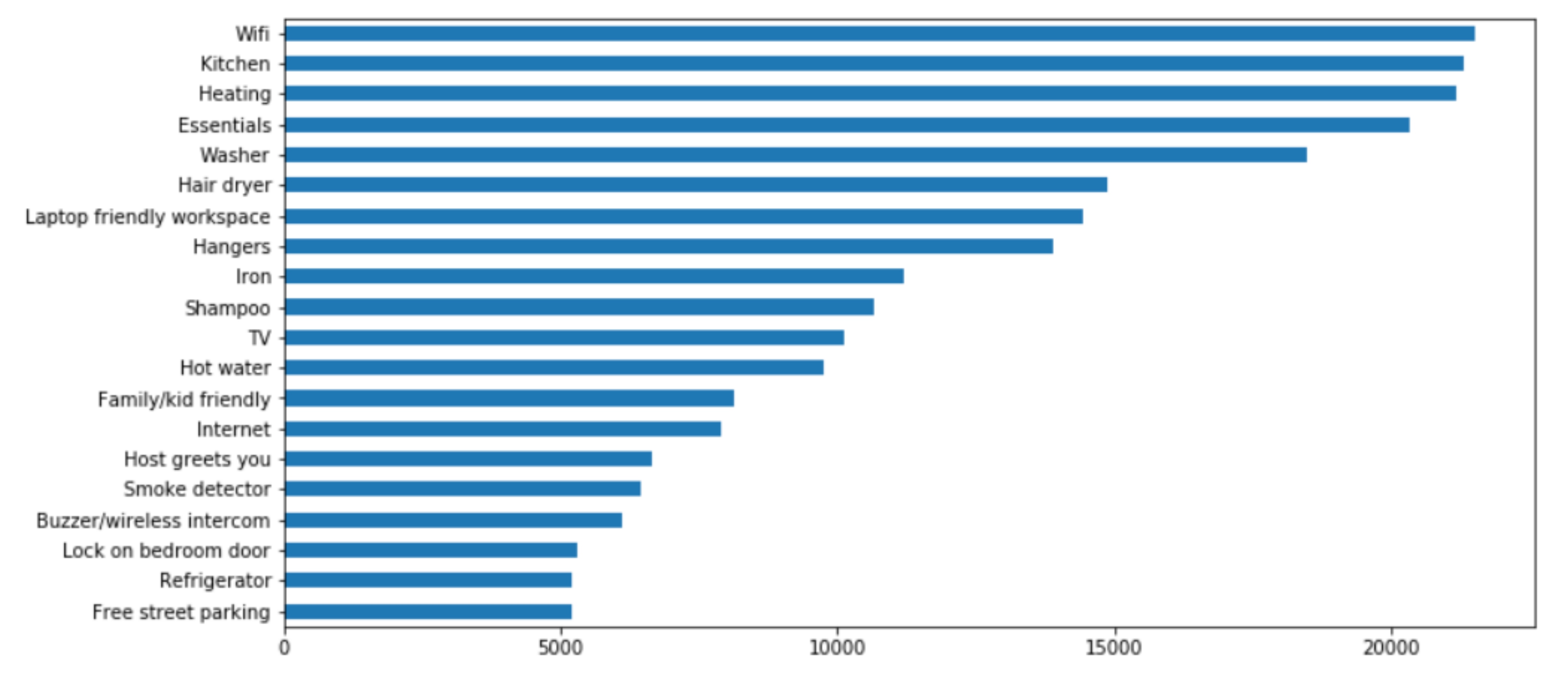


```
{TV,"Cable TV",Wifi,Kitchen,Gym,Heating,"Family/kid friendly","Smoke detector",Essentials,Shampoo,"Lock on bedroom
door",Hangers,"Hair dryer",Iron,"Laptop friendly workspace","Private living room",Bathtub,"Hot water","Bed
linens","Extra pillows and blankets",Microwave,"Coffee maker",Refrigerator,Dishwasher,"Dishes and
silverware","Cooking basics",Stove,"Luggage dropoff allowed","Long term stays allowed"}
```

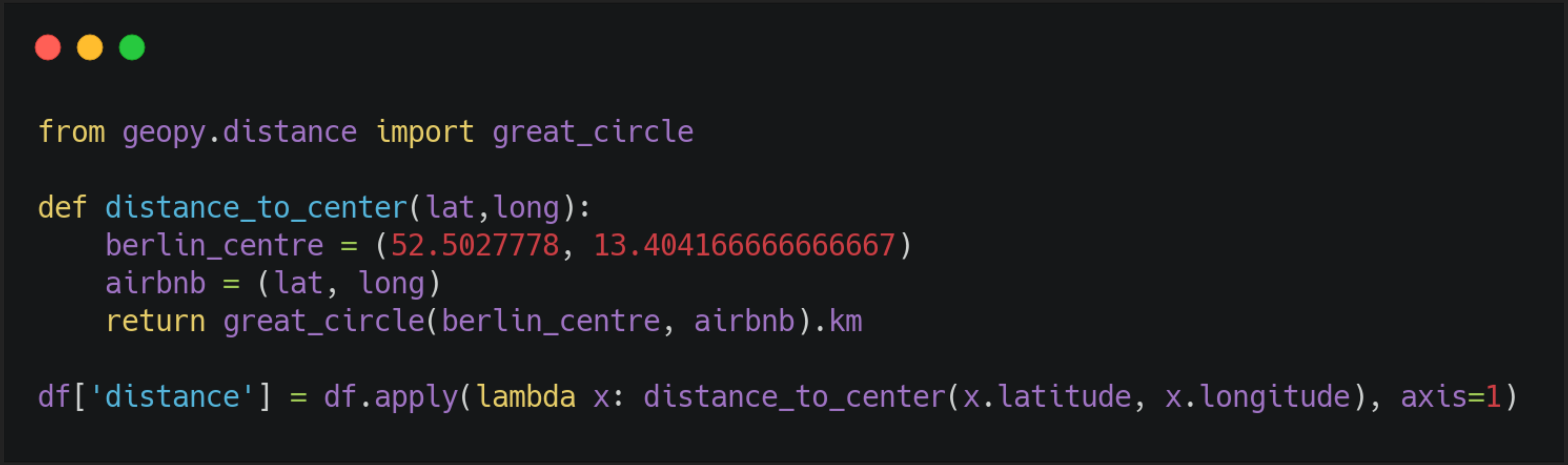


```
listing['amenities'] = listing['amenities'].str.strip('{}')\
                                            .str.replace('"', "")\
                                            .str.split(',')
```

AMENITIES – REPARTITION



DISTANCE TO BERLIN CENTER

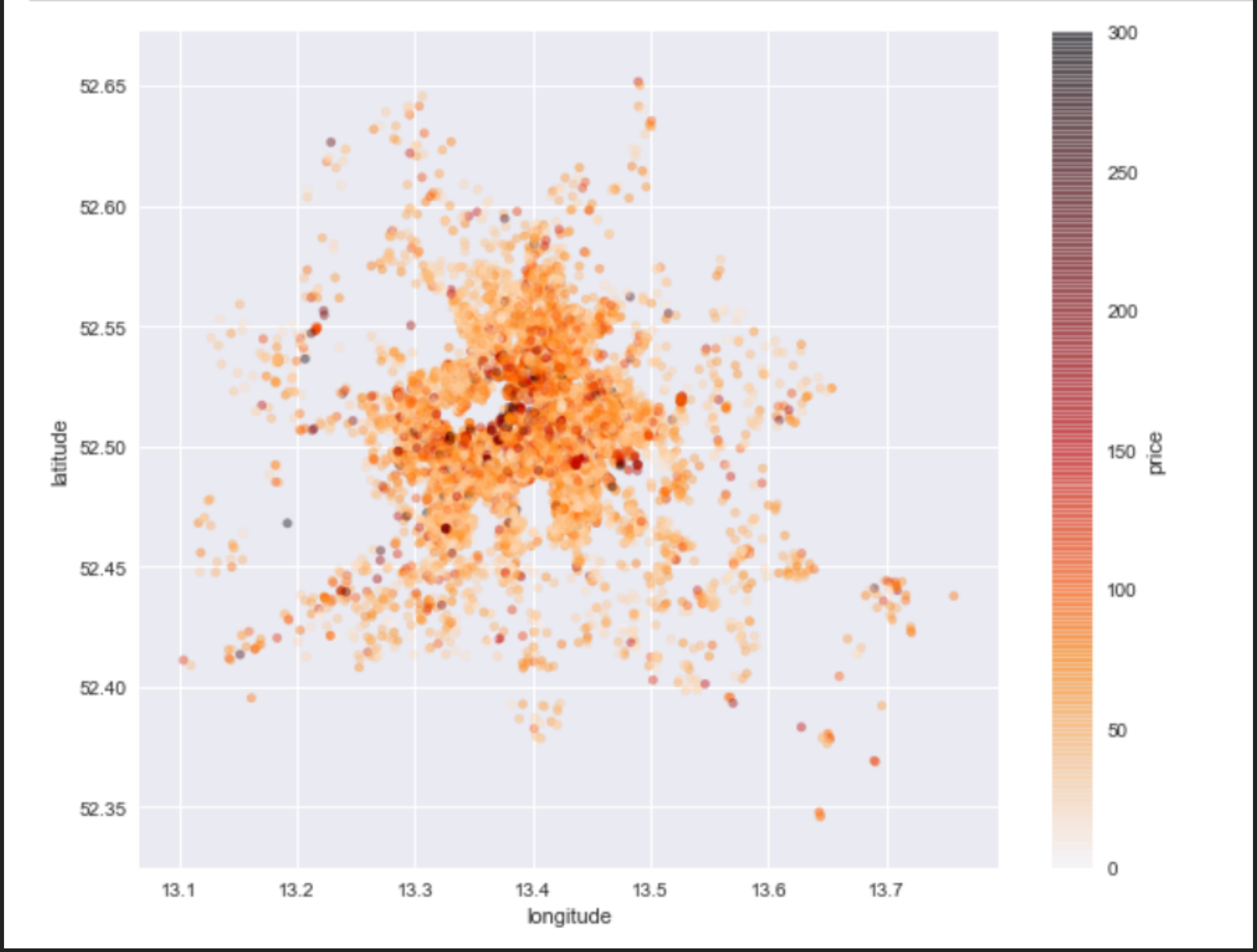


```
from geopy.distance import great_circle

def distance_to_center(lat,long):
    berlin_centre = (52.5027778, 13.404166666666667)
    airbnb = (lat, long)
    return great_circle(berlin_centre, airbnb).km

df['distance'] = df.apply(lambda x: distance_to_center(x.latitude, x.longitude), axis=1)
```

DISTANCE TO BERLIN CENTER



PRICE PREDICTION

- Train/Test Split
- Standardization
- RandomForest
  - Mean Absolute Error: 22.72 $
  - Score: 28.03 %
- XGBoost
  - Mean Absolute Error: 21.25 $
  - Score: 35.2 %

PREDICTING PRICE

CONCLUSION