Exploratory Data Analysis for Machine Learning

Final Peer Assignment

Yulia Shutko

Table of content:

Introduction

In this project ranking of fast-food restaurants in the USA, the number of restaurants per capita and area in the top 30 cities, as well as which fast food restaurants are popular in these cities have been explored.

For this project two data sets have been used, namely “Fast Food Restaurants USA” from the website Kaggle.com (link to the data-set: https://www.kaggle.com/khushishahh/fast-food-restaurants-across-us) and data set of biggest cities population in the USA from the website worldpopulationreview.com (link to web-page https://worldpopulationreview.com/us-cities). The population data set have been obtained with the help of web scraping technic.

Fast food restaurants data set-contains 10000 rows and 11 following columns: “Unnamed: 0”, “Address” “Categories”, “City”, “Country”, “Latitude”, “Longitude”, “Name”, “Postal Code”, “Province”, “Websites”. For further analysis columns “Unnamed 0” and “Website” have been not used. In this data set, it is important to use columns “City” and “Province” (state) together to avoid wrong data for cities with the same names but different locations (e.g. there are 34 cities named Springfield in different states).

Population data set has 200 entries and 8 columns: “Rank”, “Name”, “State”, “2021 Pop.”, “2010 Census”, “Change”, “Density (mi²)”, “Area (mi²)”.

In this project exploratory data analysis, feature engineering and significance tests have been performed.

EDA: data preparation

The initial plan for EDA for two data set includes:

1. Determine which fast food restaurant chains are most popular.
2. Determine the top 30 cities with the highest number of restaurants.
3. Determine the 10 most popular restaurants in these cities.
4. Determine the number of restaurants per capita for the top 30 cities.

The first step of data cleaning for the population data set is to gather initial information with methods *.shape*, *.dtypes*, *.info()* and *.columns*. This basic action shows how big is our data set, which types of data columns contain, whether there are missed values or not. In this case, the data frame has 200 rows and 8 columns, has no issues with missed values and incorrect data types. But for further data analysis columns “2010 Census” and “Change” have been dropped. As far as in this project 2 different data set have been used, it was necessary to have identical designations for states in both data frames, therefore full names were renamed to their abbreviations. For this goal a dictionary with full names and abbreviations for all states was created, then each entry was renamed with the *.map* method. Also features with sq.miles (“Density of population” and “Area”) were recalculated into a metric system (sq. km).

As well as for the population data set, the first step in data cleaning for the restaurant data set was gathering initial information with mentioned above methods. This data frame contains 10000 rows and 11 columns, has no issues with missed values or incorrect data types. Three columns "Unnamed: 0", "Websites" and "Country" were dropped. For this data set, it is important to make all entries uniform, because some restaurants’ names are spelt in different ways,e.g. Macdonalds and Macdonald’s or SONIC DRIVE -IN and Sonic Drive In. Methods *.str.replace()* and *.str.title()* solved that problem. Another problem in this data set is that New York City is designated as three separate cities: New York, Brooklyn and Bronx. This situation was solved with the help of the renaming of Bronx and Brooklyn into New York.

For the first point in the initial EDA plan method .value_counts() was applied for the “Name” column to determine which fast-food chains are most popular.

For the second point in the initial EDA plan, method *.groupby().count()* by city and state was applied. To determine the top 30 cities *.sort_values* by a total number of restaurants in each city was used.

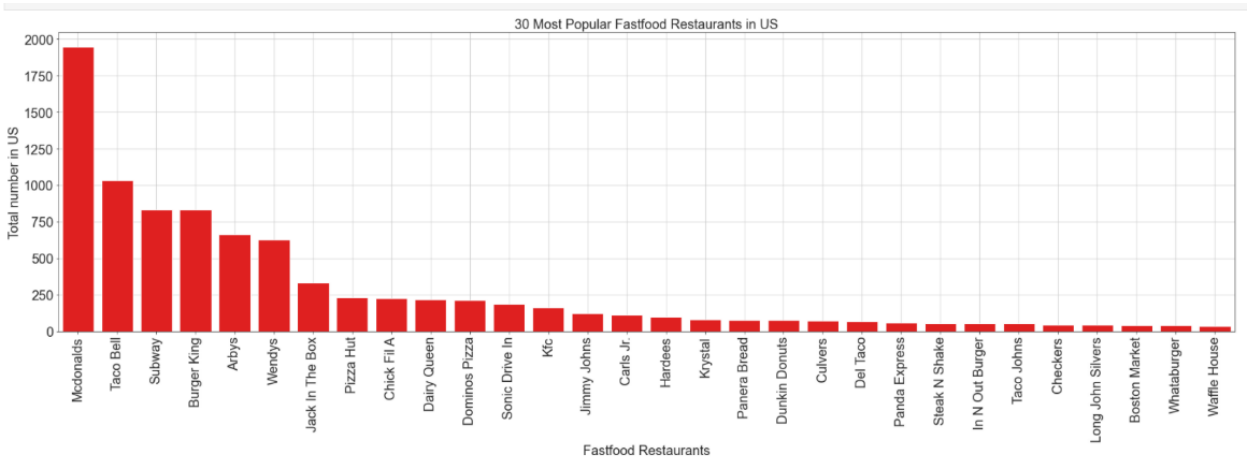
For the third point in the EDA plan, one-hot encoding of each restaurant was used. Then all restaurants were grouped by the city in the top 30, then the top 10 chains were determined.

For the fourth point in EDA, two data sets were merged by city and state. Also, columns “Number of restaurants per capita”, “Number of restaurants per 100,000 people”, “Number of restaurants per sq. km” and “Number of Restaurants per 100 sq. km” were calculated.

Key findings and insights

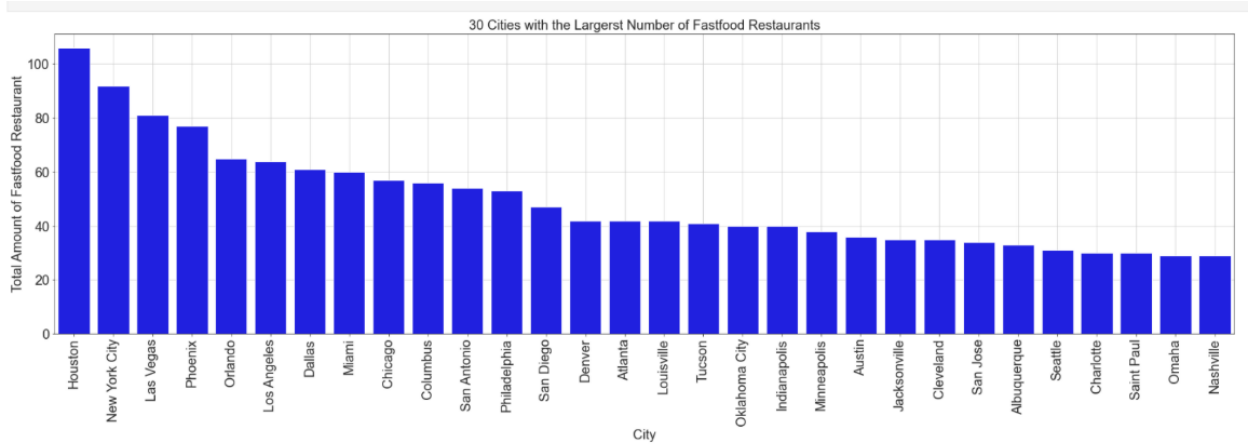
In the “Fast food restaurants in the US” data set there are 544 unique names. But here only the top 30 popular restaurants are considered.

In picture 1, 30 most popular fast food restaurant chains across the USA are represented. The leading chain is McDonald’s with 1948 restaurants across the USA. Other 3 leaders: Taco Bell, Subway, Burger King, Arby’s and Wendy’s have 2 times low numbers 1032, 833, 833, respectively. Arby’s and Wendy’ 666 and 628.



Picture 1. Top 30 most popular fast-food restaurants across US

In picture 2, 30 cities with the highest number of fast-food restaurants are represented.



Picture 2. Top 30 cities with the highest numbers of fast-food restaurants

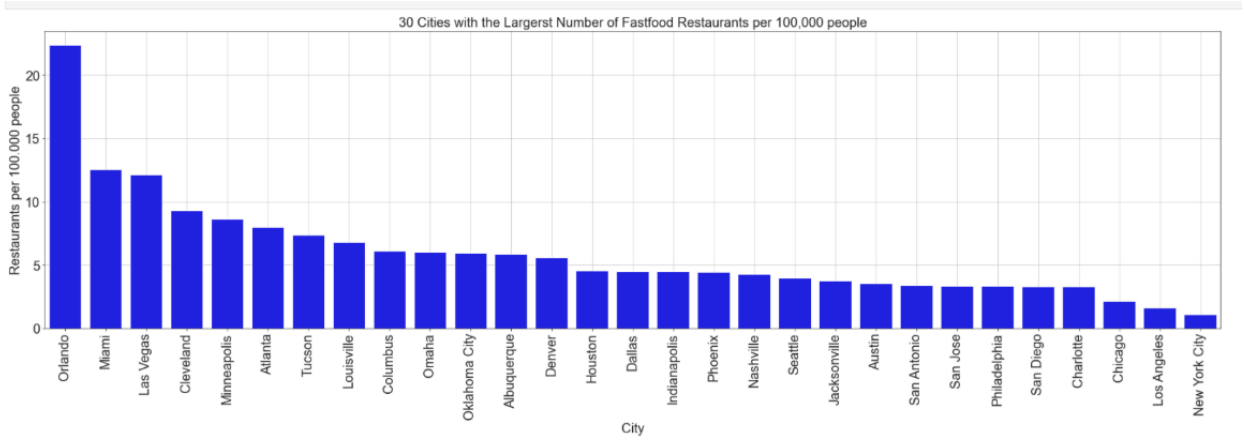
As it can be seen from the bar plot above, the leading city is Houston, but not New York as the biggest city in the USA.

	City	Rank	State	2021 Pop.	Population Density (sq.km)	Area (sq.km)	Total Amount of Fastfood Restaurant	Restaurants per capita	Restaurants per 100.000 people
0	Houston	4.0	TX	2323660.0	1452.0	1600.475	106	0.000046	4.561769
1	New York City	1.0	NY	8230290.0	10960.0	750.950	92	0.000011	1.117822
2	Las Vegas	25.0	NV	667501.0	1883.2	354.425	81	0.000121	12.134813
3	Phoenix	5.0	AZ	1733630.0	1339.6	1294.175	77	0.000044	4.441548
4	Orlando	71.0	FL	290520.0	1051.2	276.400	65	0.000224	22.373675
5	Los Angeles	2.0	CA	3983540.0	3397.6	1172.400	64	0.000016	1.606611
6	Dallas	9.0	TX	1347120.0	1586.0	849.350	61	0.000045	4.528179
7	Miami	42.0	FL	478251.0	5314.4	90.000	60	0.000125	12.545713
8	Chicago	3.0	IL	2679060.0	4713.2	568.425	57	0.000021	2.127596
9	Columbus	14.0	OH	913921.0	1667.6	548.000	56	0.000061	6.127444
10	San Antonio	7.0	TX	1581730.0	1304.4	1212.775	54	0.000034	3.413983
11	Philadelphia	6.0	PA	1585010.0	4721.6	335.700	53	0.000033	3.343827
12	San Diego	8.0	CA	1427720.0	1752.4	814.700	47	0.000033	3.291962
13	Denver	19.0	CO	749103.0	1954.8	383.225	42	0.000056	5.606706
14	Atlanta	37.0	GA	524067.0	1544.4	339.350	42	0.000080	8.014242
15	Louisville	29.0	KY	615924.0	935.2	658.575	42	0.000068	6.819023
16	Tucson	33.0	AZ	554503.0	932.0	595.025	41	0.000074	7.394009
17	Oklahoma City	24.0	OK	669347.0	441.6	1516.125	40	0.000060	5.975974
18	Indianapolis	16.0	IN	887232.0	981.6	903.925	40	0.000045	4.508404
19	Minneapolis	46.0	MN	439012.0	3252.0	135.000	38	0.000087	8.655800
20	Austin	10.0	TX	1011790.0	1264.8	799.850	36	0.000036	3.558051
21	Jacksonville	13.0	FL	929647.0	497.6	1868.675	35	0.000038	3.764870
22	Cleveland	54.0	OH	376599.0	1938.8	194.225	35	0.000093	9.293705
23	San Jose	11.0	CA	1009340.0	2270.8	444.525	34	0.000034	3.368538
24	Albuquerque	32.0	NM	562281.0	1201.2	468.050	33	0.000059	5.868952
25	Seattle	18.0	WA	776555.0	3704.0	209.650	31	0.000040	3.991990
26	Charlotte	15.0	NC	912096.0	1187.6	768.100	30	0.000033	3.289127
27	Omaha	41.0	NE	479978.0	1362.0	352.450	29	0.000060	6.041944
28	Nashville	23.0	TN	678448.0	570.8	1188.850	29	0.000043	4.274462

Table 1. Top 30 restaurants with the highest number of restaurants.

On the third position is Las Vegas, but it has 25th place in the population chart. According to the number of restaurants and population (see Table 1), it is possible to say that there is a moderate positive correlation between these two variables.

In picture 3, distribution of the number of restaurants per 100,000 people in the Top 30 cities. Here 3 biggest cities (NYC, Los Angeles and Chicago) are at the end of the chart, but smaller cities have a higher density of restaurants. This is explained by the big population and a rather low number of fast-food restaurants, while smaller cities have more restaurants per 100,000 people of population.



Picture 3. Distribution of number of restaurants per 100,000 people in Top 30 cities

Hypothesis testing and significance test

Null Hypothesis #1: The number of restaurants does not depend on the area of cities.

Alternative Hypothesis #1: The number of restaurants depends on the city's area.

Null Hypothesis #2: There are more fast-food restaurants in eastern states

Alternative Hypothesis #2: The number of restaurants is higher in other parts of the USA

Hypotheses for testing:

Null Hypothesis #3: The number of restaurants does not depend on population.

Alternative Hypothesis #3: This number depends on the city's population.

For hypothesis testing, it is necessary to get new data set out "Fast-food Restaurants across the US" and population data. In this case, two data set were merged by city and state, to avoid incorrect data.

Citation from https://stackabuse.com/statistical-hypothesis-analysis-in-python-with-anovas-chi-square-and-pearson-correlation/:

"The Pearson Correlation test is used to analyze the strength of a relationship between two provided variables, both quantitative.

The value, or strength of the Pearson correlation, will be between +1 and -1. A correlation of 1 indicates a perfect association between the variables, and the correlation is either positive or negative. Correlation coefficients near 0 indicate very weak, almost non-existent, correlations. While there are other ways of measuring correlations between two variables, such as Spearman Correlation or Kendall Rank Correlation, Pearson correlation is probably the most commonly used correlational test."

For the test a Pearson's correlation test was used, significance level set to $\alpha = 0{,}05$.

```
[62]: from scipy import stats
      x = hyp_samp["Total Amount of Fastfood Restaurant"]
      y = hyp_samp ["2021 Pop."]

      pearson_coef, p_value = stats.pearsonr(x,y)
      print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)

      The Pearson Correlation Coefficient is 0.7347049379381756  with a P-value of P = 1.2041496032656178e-31
```

From the results of this test, the null hypothesis should be rejected, and alternative accepted. Here we see quite strong positive linear (0,734) and statistically significant (p-value <0.001) correlations.

Further analysis

For the represented data set it is possible to execute the following analysis:

1. Number of restaurants per state
2. Top 20 restaurants in states
3. Fast food restaurants per capita for all states
4. The density of restaurants in states per 100sq.rm

Summary

In this project data set from different sources, namely kaggle.com and worldpopulationreview.com, were obtained. These data sets were cleaned and processed for further EDA analysis. Key findings and insights were represented. Three hypotheses were described and one significance test was executed.