Exploratory Analysis of Electoral Participation Data Set for the 2021 Parliamentary Elections in Catalonia

Elitza Maneva

1. Description of the data set

This data set was created for the Capstone Project of the Data Science IBM Specialization on Coursera. The goal is to look for correlations between election participation and neighborhood characteristics.

There are two main sources of data. The first source is the publicly available data set from the elections held on February 14, 2021 in Catalonia. In particular we use the census and voter participation by electoral sections. The second source of data is the Foursquare API which we queried for venues in the proximity of the electoral sections. In order to create the data set we also had to use geographic data about the electoral sections in Catalonia, which is publicly available on a government website.

Before querying the Foursquare API we removed some outliers, namely sections with census less than 400. These generally have turnout that is much higher than average, and also correspond to geographic areas with very sparse venue data on Foursquare. After removing these sections our table has 4643 rows corresponding to the remaining electoral sections. Each section is identified by its municipality, district and section number, as well as geographic coordinates for the centroid of the area it encompasses.

The venue data was processed as part of the Capstone Project. The venue categories were reduced to 30 types (down from 698 that were provided by Foursquare). We have columns for the number of venues returned by Foursquare for each of these 30 categories. The total number of venues is 175,354. The average number of venues per electoral section is 109 (note that some venues correspond to more than one electoral section, especially in the very dense areas).

Additionally there are columns for the census, for the turnout, and for the population density.

The total number of citizens in the census is 5,368,992, and the number who voted is 2,874,610 for a total turnout of 53.54%. The country occupies an area of 32,108 km^2.

On the map below you can see the geographic locations of the electoral sections. The transparent points are the ones that were removed from the set for having too few names in the census. The red dots correspond to sections with relatively low turnout, and the blue dots to relatively high turnout.

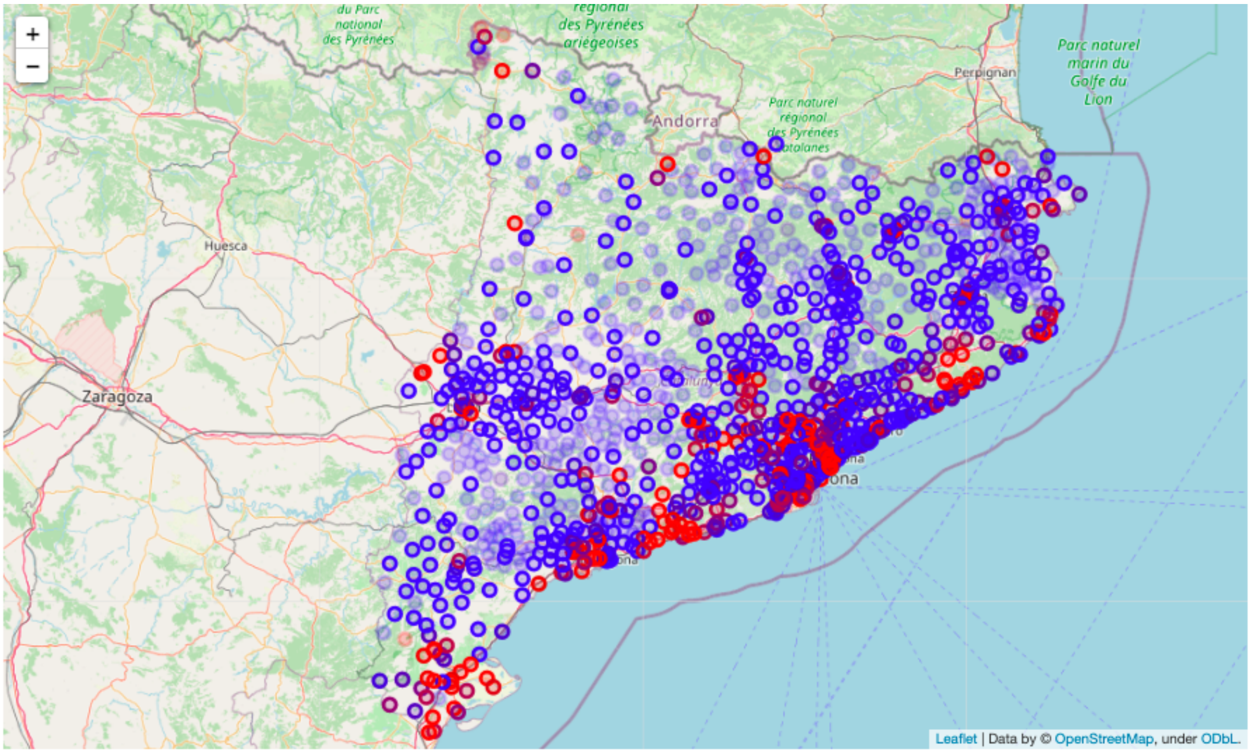
+
−
du Parc national des Pyrénées
régional des Pyrénées ariégeoises
Andorra
Parc naturel régional des Pyrénées catalanes
Perpignan
Parc naturel marin du Golfe du Lion
Huesca
Zaragoza
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Electoral sections in Catalonia, color coded depending on the level of voter turnout in the 2021 elections. Red corresponds to lower turnout.

As an example, below we have a map with the locations of the venues associated with 10 different electoral districts in the municipality of Esplugues de Llobregat.

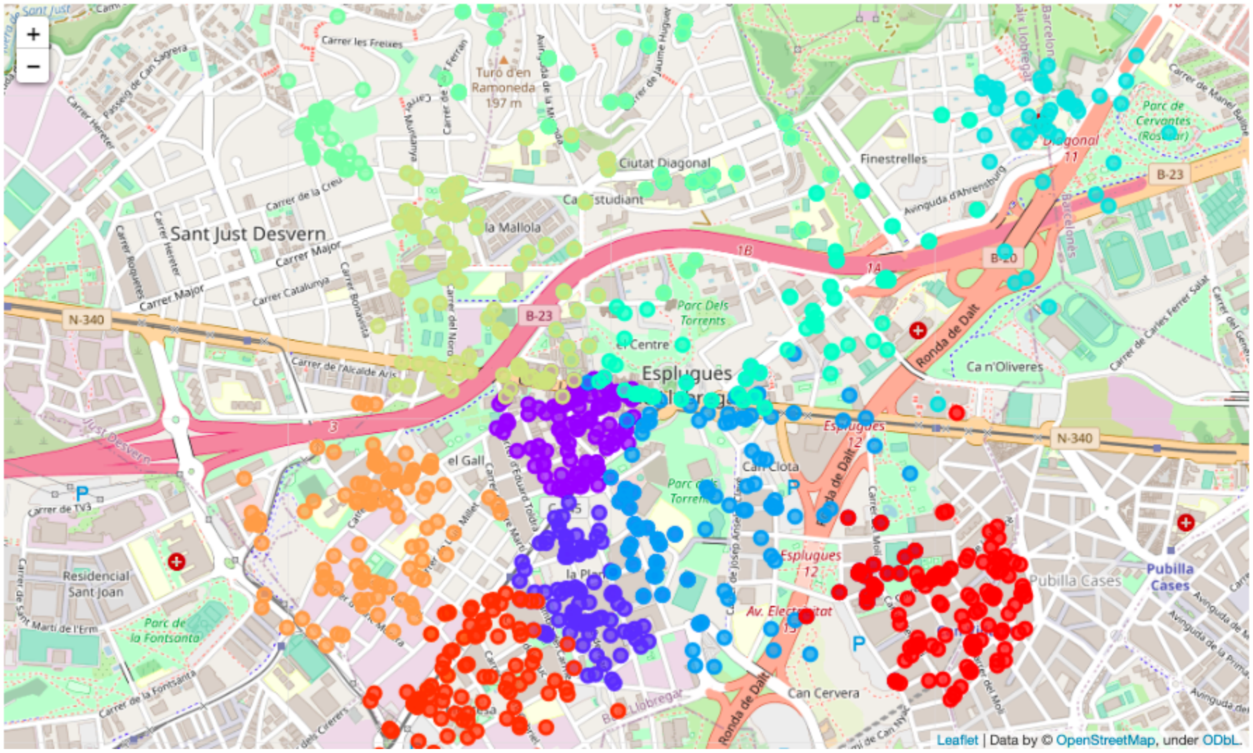

Figure 2. Locations of the venues returned by the Foursquare API for the 10 electoral districts in the municipality of Esplugues de Llobregat.

We considered querying Foursquare only for districts and not for each section, however this leads to giving much higher weight to districts with few sections. Since the sizes of the sections are close to normally distributed, it makes more sense to consider each section as a data point. Notice that on the map of Esplugues de Llobregat we show the venues that are returned for the 10 districts. When we query sections we get a lot of overlap between different sections. We have to take care of this overlap in order to assure that the observations we use are independent.

2. Plan for data exploration

As a first step we will check the distribution of our target variable - the turnout, and see whether it is normally distributed. Second, since we can already guess that population density will not be normally distributed, we will consider using the logarithm of the population density.

Having already reduced the number of category types to 30, we can proceed to look for venue types that are most correlated to turnout. We will have to consider only a subset of the sections in order to make sure that the observations are independent. In particular we need a subset of sections that do not share venues.

Once we have a set of independent observations, we will check for skew in the independent variables. We expect to observe right skew for most categories as we are dealing with geographic locations of venues, which tend to have Poisson type distributions. We will do a box-cox transformation and observe how much that changes our distributions.

3. Data cleaning and feature engineering

3.1. Distribution of voter turn-out

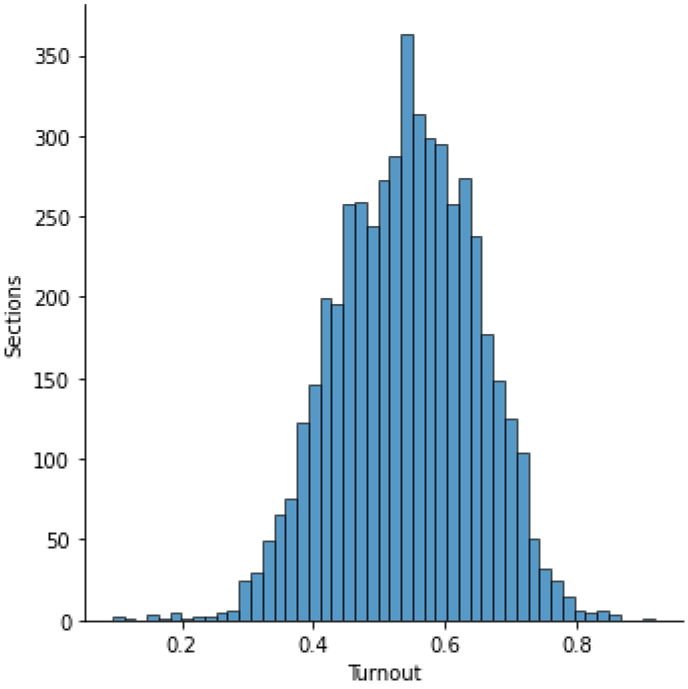

Figure 3. Distribution of voter turnout by sections

3.2. Sizes of sections (census)

As part of the data cleaning process we removed 90 sections for which we didn't have geographic coordinates. Next, we removed sections with a very small census as they can be considered outliers. The distribution of the sizes of the sections is shown in Figure 4. The small sections generally have high turnout as shown in Figure 5.

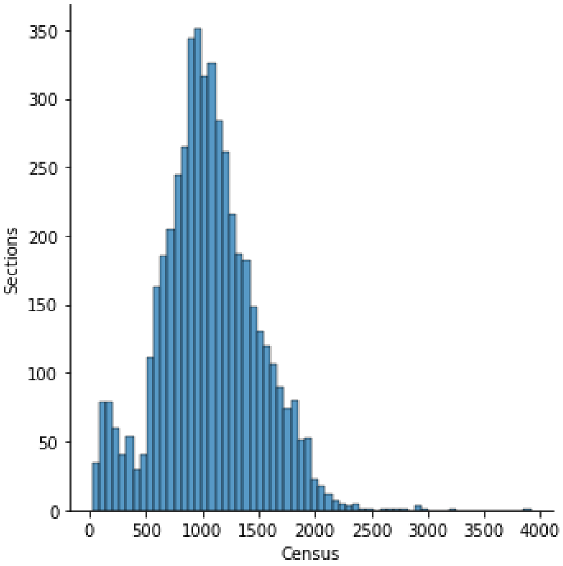

Figure 4. Distribution of the census size of the sections.

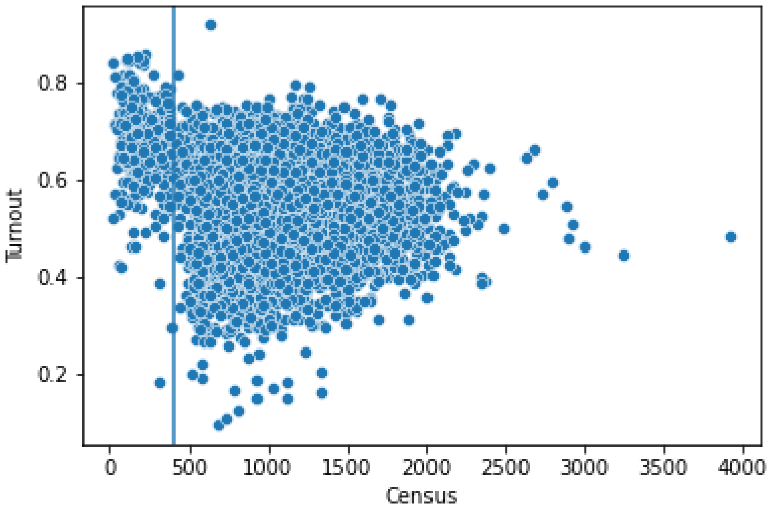

Figure 5. Voter turnout with respect to census size. The small sections that we remove from the set before querying Foursquare are the ones to the left of the vertical line.

3.3. Population density

The distribution of the logarithms of the population density and its relationship with voter turnout are plotted below.

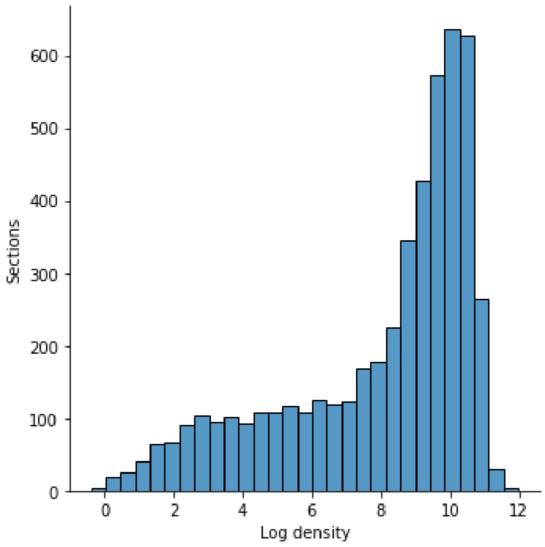

Figure 6. Distribution of the logarithm of the population density by electoral sections.

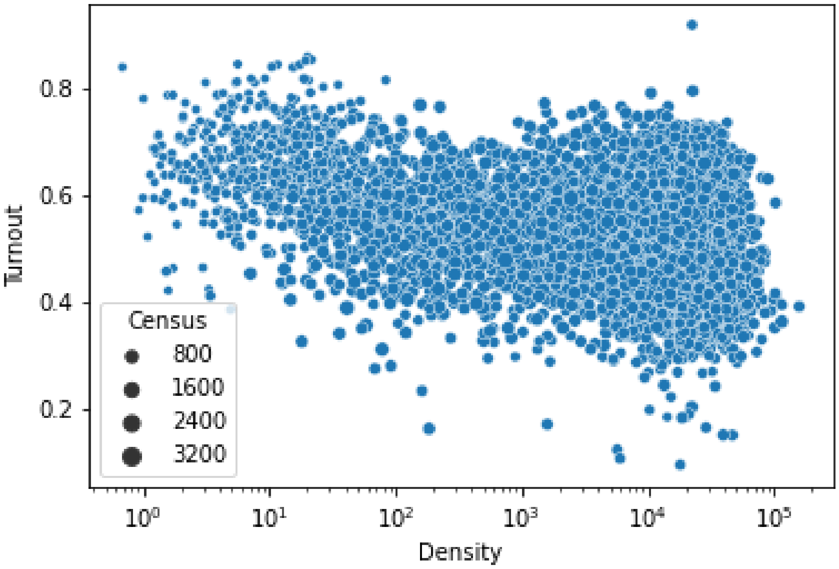

Figure 7. Turnout versus population density by electoral sections.

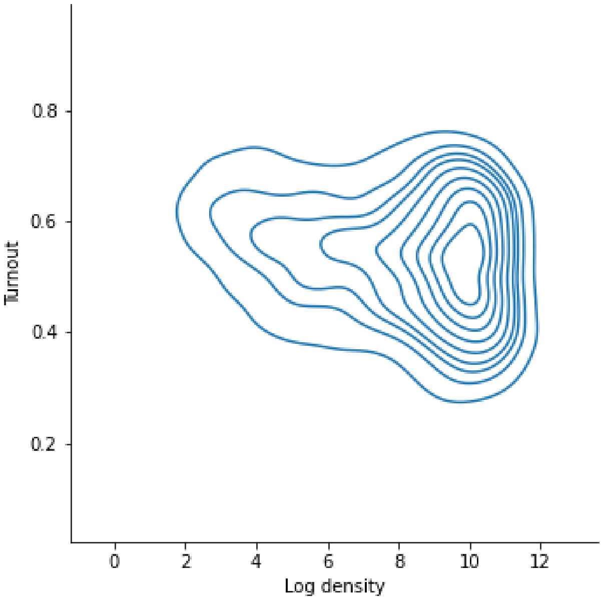

Figure 8. Distribution of the turnout versus the logarithm of the population density (only sections with more than 400 people in the census are included).

3.4. Categories of venues

We reduced the categories of venues to 30 down from 698 by merging some and by removing low frequency ones.

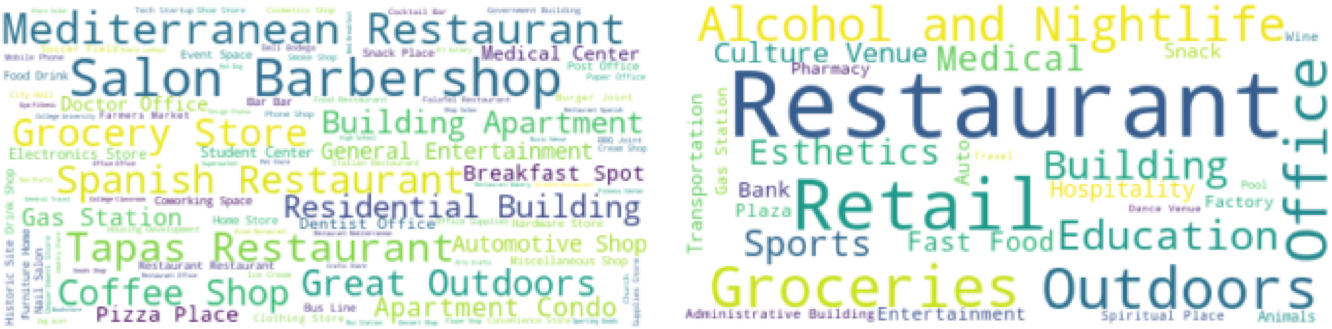

Figure 9. Word cloud of categories before and after summarizing the categories.

3.5. Selecting independent observations

We went over the electoral sections in order and only included in our final data set those that have at most 4 venues in common with the sections that were already selected. This way we avoid dependencies between the observations included in the final data set. At the end of this selection process we have 1049 electoral sections in our data set (down from 4643).

3.6. Skew of the independent variables

As expected, all of the venue categories have positive skew. The ones with least skew are the Restaurants and Grocery Stores. The highest skew by far is the Wine category.

We fix the skew by applying a box-cox transformation. In the following figure you can see the original distribution for the category of “Office”, as well as its distribution after a log(1+p) transformation is applied, and the distribution after the box-cox transformation.

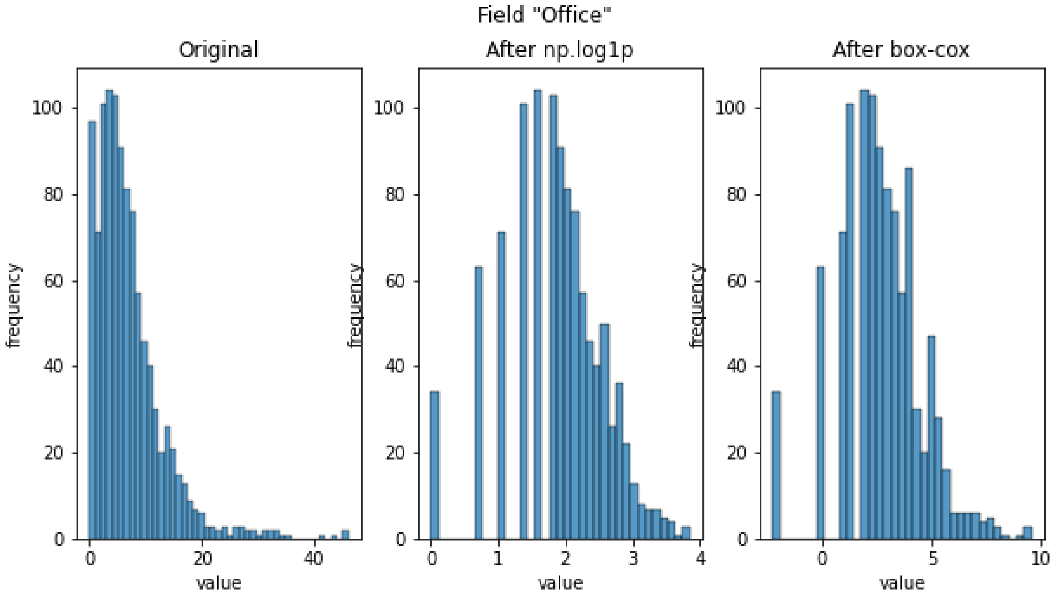

Figure 10. The distribution of the number of Office venues, as well as the log(1+p) and the box-cox transformation of this distribution.

4. Key Findings and Insights of EDA

We calculated the correlation of each venue category to the voter turnout after having applied the box-cox transformation. The values are compared in Figure A.1. It is important to keep in mind also the relative frequencies of each venue. In Figure A.2. you can see the total occurrences of each type of venue for our selection of 1047 electoral sections. The skew of the venue types is plotted in Figure A.3.

Figure A.4. shows the scatter plots (after box-cox transformations) for all pairs among the turnout and the three venue categories that are most positively correlated with voter turnout, as well as the three venue categories that are most negatively correlated with voter turnout.

Finally, we look at a pairwise correlation plot for the 15 venue categories most strongly correlated with voter turnout. The categories are sorted by their correlation with the turnout.

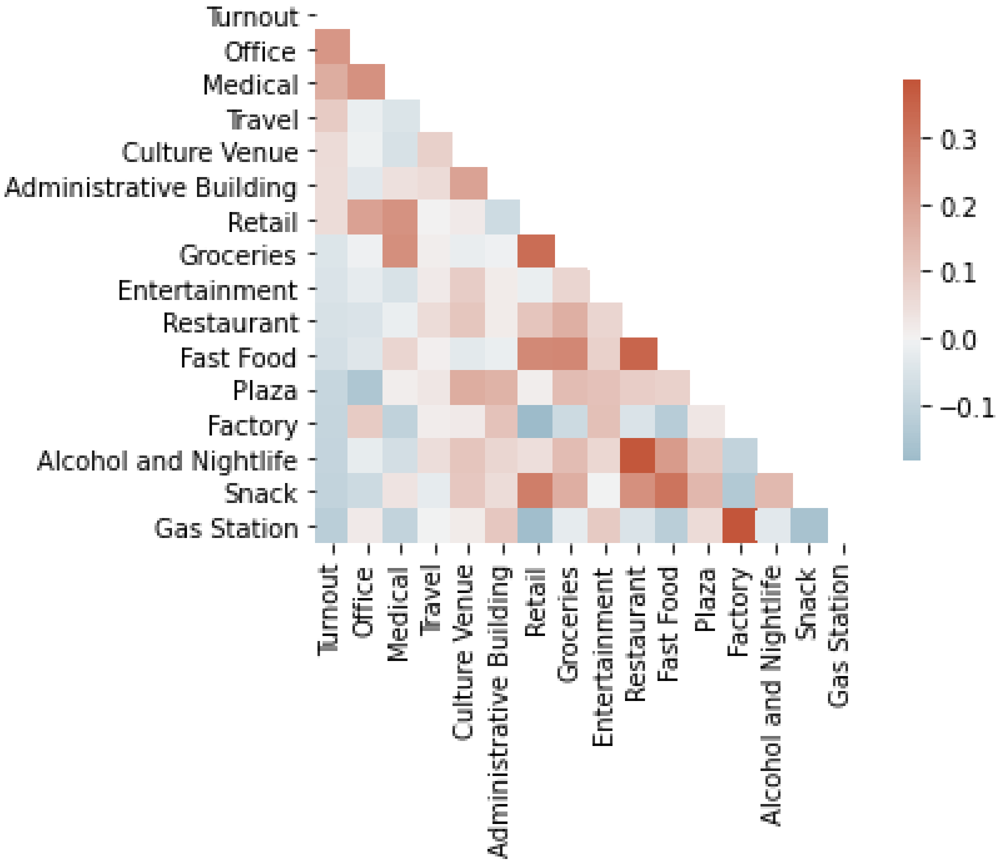

Figure 11. Heatmap of the correlations between the top 15 categories most correlated with voter turnout.

5. Some Initial Hypothesis

We tested the null hypothesis that voter turnout is normally distributed. The distribution of voter turnout by electoral sections is shown in Figure 3. We performed the D'Agostino K^2 test, with criterion for acceptance of p>0.05. We obtained p-value 0.0004, therefore we can say that the voter turnout is not normally distributed.

We tested the same null hypothesis for the 10 largest municipalities and it was not rejected for half of them - Badalona, Terrassa, Tarragona, Santa Coloma de Gramenet, and Reus. On the other hand it was rejected for the biggest municipality, that of Barcelona.

We have calculated the Pearson correlation between voter turnout and the different venue categories as shown in Figure 11. However, we still need to test the significance of the correlation coefficients for our sample size. Given that our sample size is relatively large (around 1000 observations) the null hypothesis of the absence of correlations is easily rejected for the venue categories with Pearson coefficient above 0.05.

Data points that correspond to geographically close locations are expected to share various characteristics. We want to test the hypothesis that even sections that are not geographically close are clustered into different types according to the frequencies of venues. We can test the hypothesis that there are clusters in our data using a Silhouette, Davies-Bouldin or Calinski-Harabasz test.

6. Suggestions for next steps

As a next step we suggest building a regression model for interpretative analysis. In our setup we are not aiming for prediction, instead, we want to find which features are most influential for voter turnout.

Furthermore, we suggest clustering the sections according to both population density and venue characteristics, with the goal to identify rural versus urban areas. We suspect that if we build a different linear regression model for each of the two clusters separately we would find coefficients that are easier to interpret.

7. Data quality

Unfortunately, we have no control over how the venues are chosen by Foursquare. We can only hope that the venues returned are a good sample of the businesses in each area.

We noticed some misclassification in the categories of the business venues. It may be possible to use the names of the venues together with a Catalan-English dictionary in order to infer their true categories.

Another way to improve the data is to query Foursquare specifically for particular types of venues for which we have some indication that are correlated with voter turnout. Currently we may be looking at a lot of data that is largely irrelevant to the target variable and we may be looking for patterns where there are few.

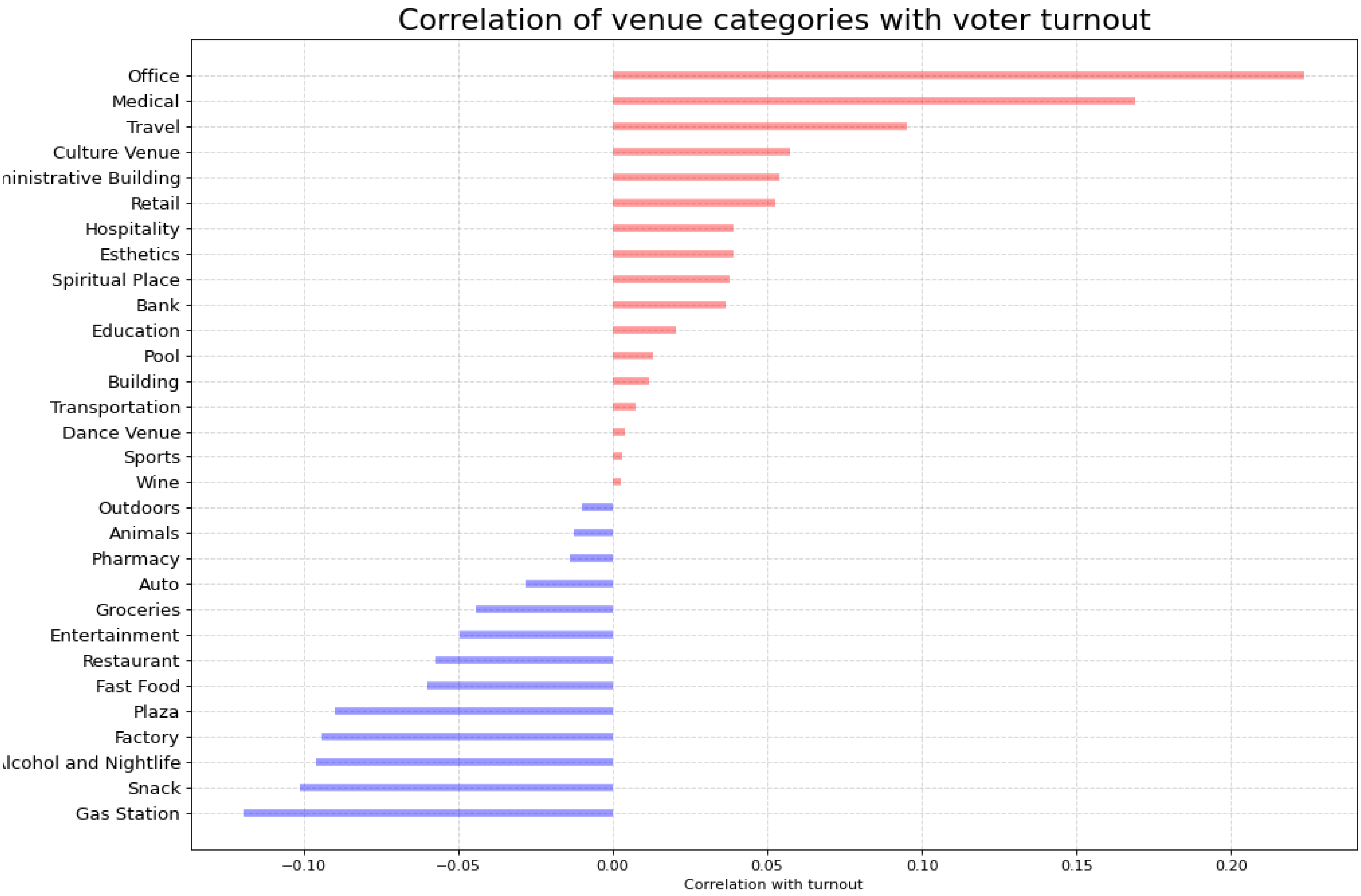

Figure A.1 Correlation of venue categories with voter turnout.

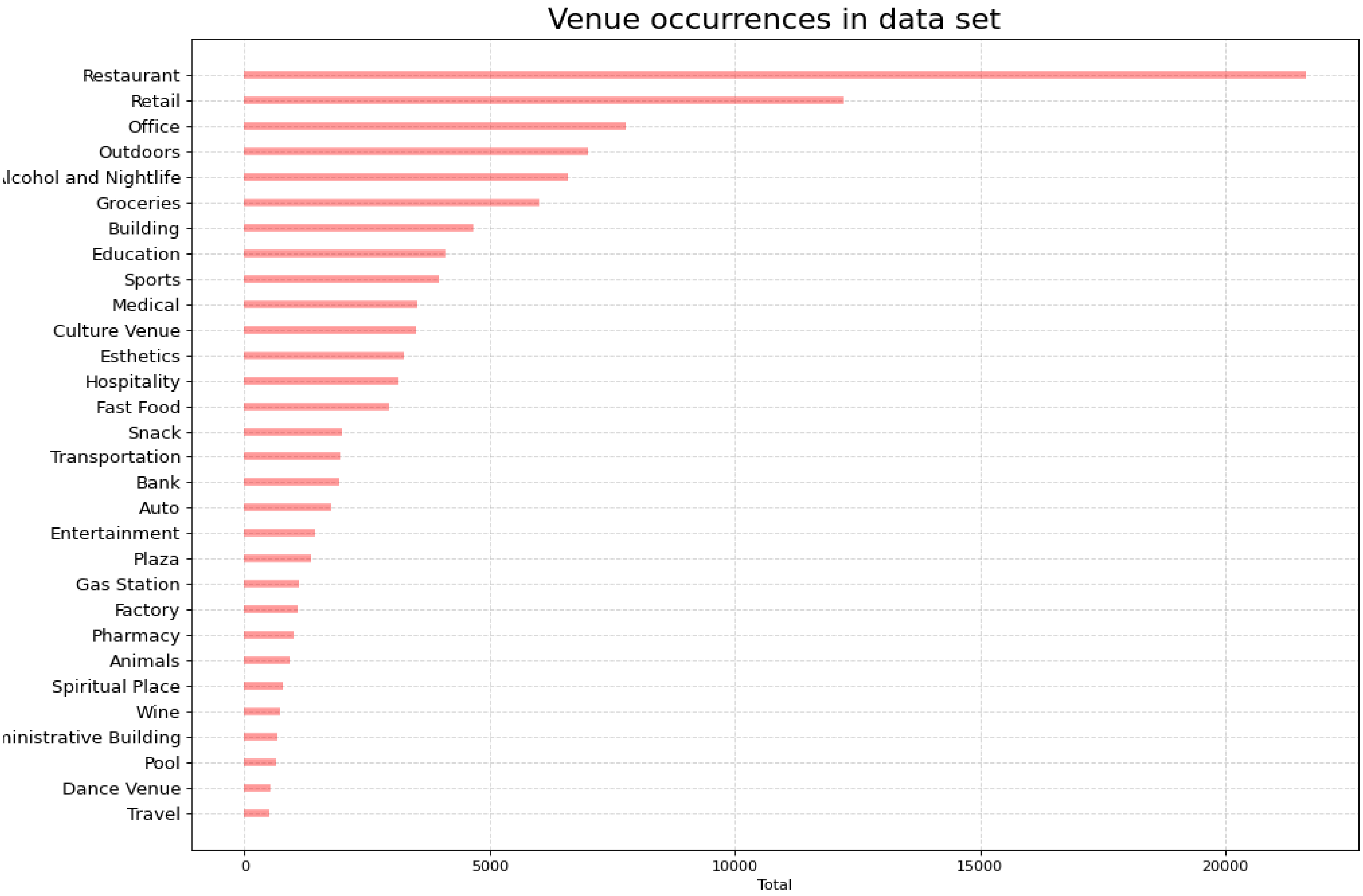

Figure A.2. Total number of venues of each kind for the 1047 selected electoral sections.

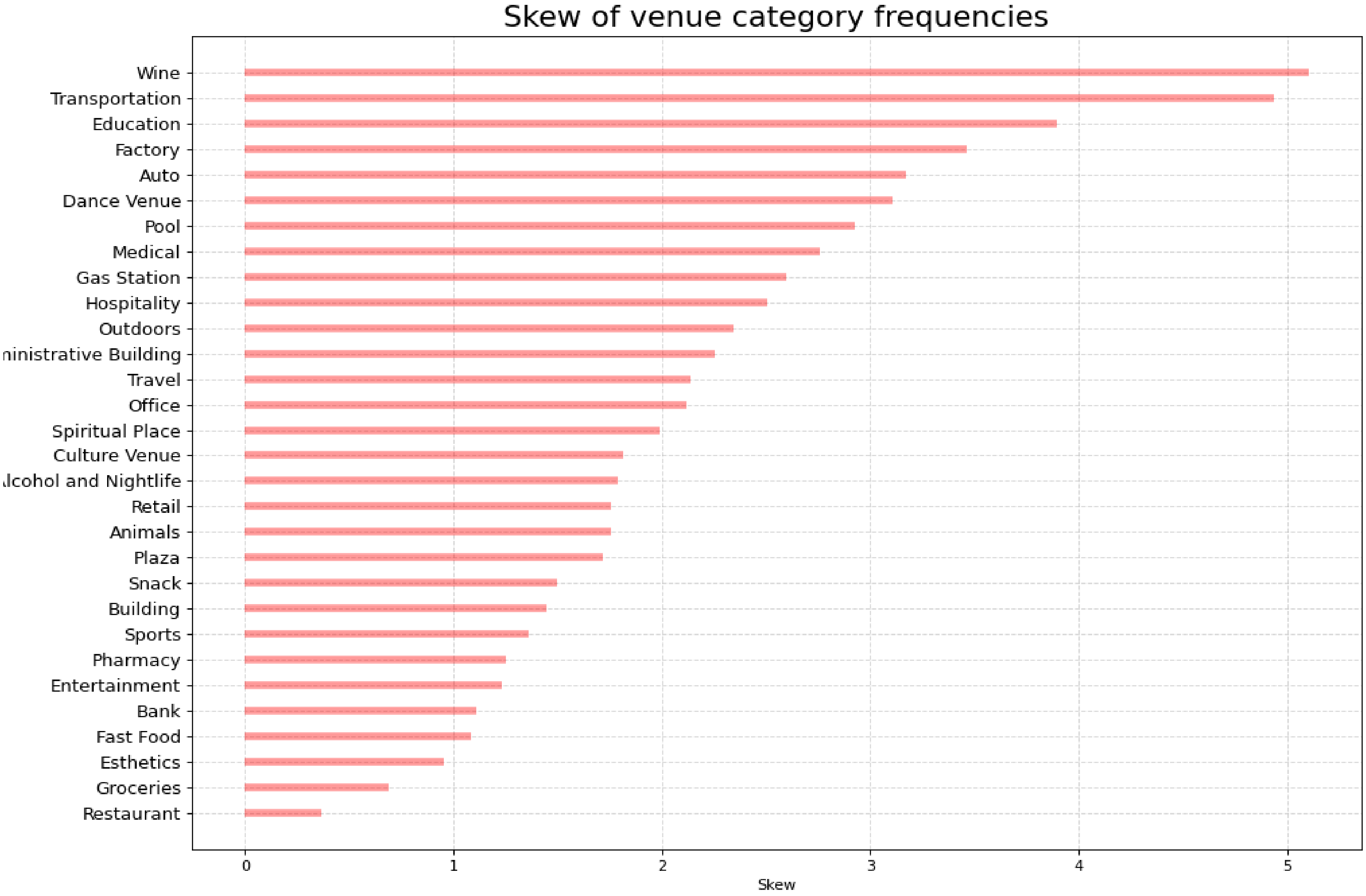

Figure A.3. Skew of the 30 venue categories.

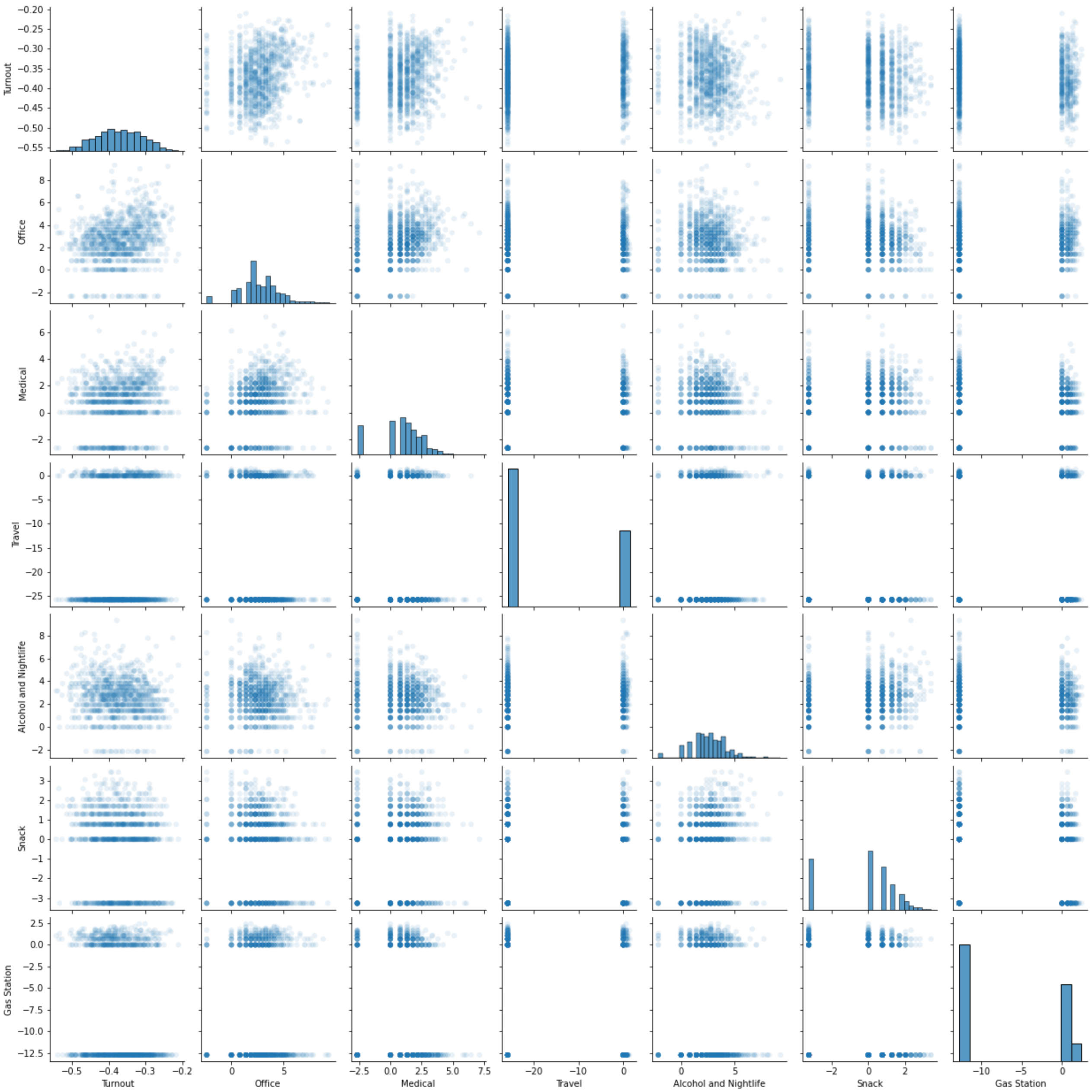

Figure A.4. Pairplot of voter turnout and 6 venue categories after box-cox transformation.