SMM635 Final Course Project

Simone Santoni

2024-11-22

Synopsis This notebook illustrates two alternative final course projects for the SMM635 course.

1 Overview of the Projects

Students are required to choose one of the two projects in Table 1. Section 2 and Section 3, describe the individual projects' context, data, and problems. Finally, Section 4 describes the materials that students are expected to submit, irrespective of the chosen project.

Table 1: Overview and key features of the two projects

Project	Title	Data domains
1	Equal employment opportunity	Geospatial & time-series
2	Platform economy	Time series

2 Project 1: Equal Employment Opportunity

2.1 Contextual information

This project's dataset contains a list of Industrial and Commercial Abatement Program (ICAP)[1] businesses that have successfully complied with Equal Employment Opportunity (EO50) requirements and have received EO50 certificate of approval from the NYC Department of Small Business Services (SBS). Executive Order No. 50 was established in 1980 during Mayor Koch's tenure. It mandates that all contractors adhere to equal employment principles, explicitly prohibiting discriminatory hiring practices. This order safeguards individuals from discrimination based on ten protected categories: race, sex, creed, age, color, disability, national origin, marital status, sexual orientation, and citizenship status.

[1]The Industrial and Commercial Abatement Program (ICAP) is a tax incentive program in New York City designed to encourage businesses to make capital investments in industrial and commercial properties. Under ICAP, eligible businesses that renovate, construct, or expand commercial and industrial buildings can receive partial property tax exemptions for a specified period. This program aims to stimulate economic development, create jobs, and modernize the city's infrastructure.

2.2 Data description

The dataset is stored in a CSV file named `e050.csv` and contains 992 cases—i.e., businesses that have received the EO50 certificate—and 16 variables, including the timestamp of the EO50 approval date and location data. Below is a sample record from the dataset.

```
import os
import pandas as pd
df1 = pd.read_csv("data/e050.csv")
df1.head(1).T
```

	0
EO50 Approval Date	06/02/2021 12:00:00 AM
Business Name	122 Fifth Associates LLC
Business Address	120 Fifth Avenue
Business City	New York
Business State	New York
Business Phone	646-630-8609
Postcode	10011
Borough	MANHATTAN
Community Board	105.0
Latitude	40.738077
Longitude	–73.992123
Council District	3.0
BIN	1015419.0
BBL	1008190037.0
Census Tract (2020)	54.0
Neighborhood Tabulation Area (NTA) (2020)	MN0501

The following bullet point list describes the individual variables:

- **EO50 Approval Date**: the date when the Business received EO50 certificate of approval from SBS (timestamp)
- **Business Name**: Business name of vendor recruiting Minority and/or Women-Owned Business Enterprise subcontractors (Text)
- **Business Address**: Address(es) of project site (text)
- **Business City**: City of project site
- **Business State**: State

- **Business Phone**: Phone number of primary contact of vendor recruiting Minority and/or Women-Owned Business Enterprise subcontractors
- **Postcode**: Zip code of site address(es)
- **Borough**: Borough in which the site is located>
- **Community Board**: Community Board number
- **Latitude**: Latitude of the site's location
- **Longitude**: Longitude of the site's location
- **Council District**: The Council District field indicates the New York City Council District where the site is located
- **BIN**: The BIN (site Identification Number) is a unique identifier for each site in the City
- **BBL**: The BBL (Borough, Block, and Lot) is a unique identifier for each tax lot in the City
- **Census Tract (2020)**: The Census Tract (Census 2020) field indicates the U.S. Census Tract where the site is located. Please note that as part of the geocoding process, leading and trailing zeros are dropped
- **Neighborhood Tabulation Area (NTA) (2020)**: The Neighborhood Tabulation Area (Census 2020) field indicates the New York City Neighborhood area where the site is located

Figure 1 shows the distribution of EO50 certificates over time.

```
import matplotlib.pyplot as plt
import seaborn as sns

df1.loc[:, "date"] = pd.to_datetime(df1["EO50 Approval Date"])
df1["month_year"] = df1["date"].dt.to_period("M")
gr = df1.groupby("month_year")
ds = pd.DataFrame(gr.size().reset_index(name="counts"))
fig = plt.figure(figsize=(8, 4))
ax = fig.add_subplot(111)
ax.plot(ds.index, ds["counts"], color="blue")
xticklabels = []
for _ in ds["month_year"].astype(str):
    if "-06" in _:
        xticklabels.append(_)
    else:
        xticklabels.append("")
ax.set_xticks(ds.index)
ax.set_xticklabels(xticklabels, rotation=45, horizontalalignment="right")
ax.set_ylabel("Counts of EO50 certificates")
ax.grid(True, which="major", axis="y", linestyle="--", linewidth=0.5)
plt.show()
```

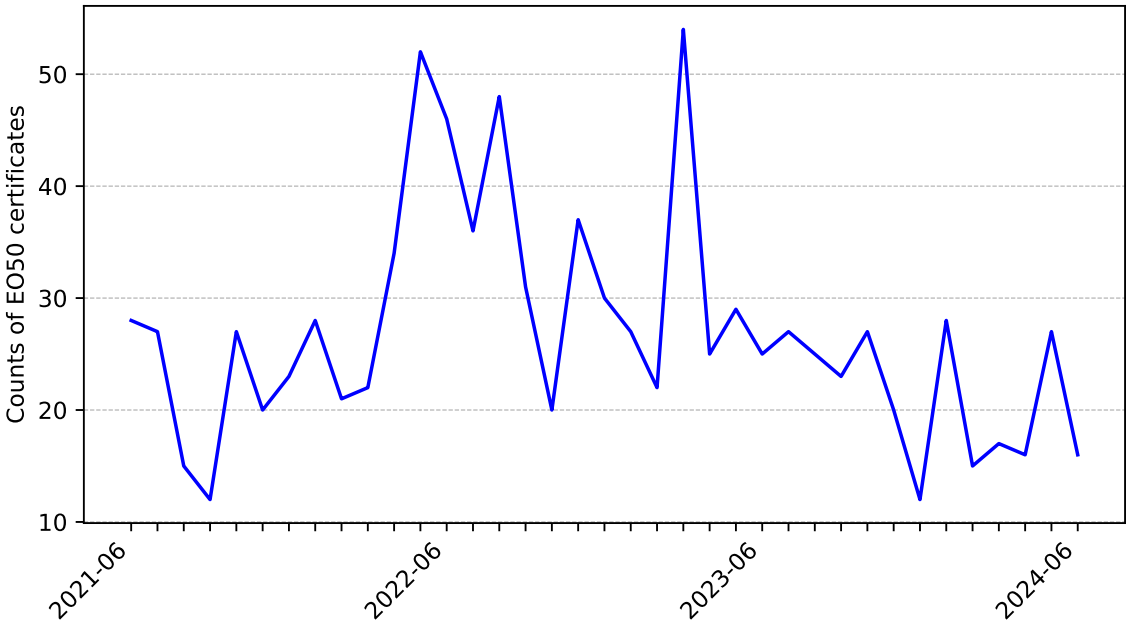

Figure 1: Distribution of EO50 certificates over time

2.3 Data visualization problem

Create two (2) plots showing the diffusion of EO50 certificates across NYC areas and over time. You may want to consider a few demographic or socio-economic indicators (e.g., borough-level indicators) to provide a more nuanced interpretation of the EO50 diffusion pattern. NYC OpenData provides a wealth of datasets that you can use to efficiently and effectively enrich the analysis. Then, discuss the main business/policy insight(s) that can be drawn from the plots you created.

3 Project 2: Platform Economy

3.1 Contextual information

The platform economy has transformed the way people work, consume, and interact with each other. Platforms like Uber, Lyft, Airbnb, and TaskRabbit have disrupted traditional industries and created new opportunities for workers and consumers. However, platform businesses may affect non-related markets in unexpected ways. This project emphasizes a regulatory change that affected the ride-sharing industry in Austin, Texas, and its impact on the local restaurant industry, and, possibly, local restaurants.

In 2016 and 2017, Uber and Lyft temporarily ceased operations in Austin, Texas, due to a dispute over local regulations governing ride-sharing services. In December 2015, Austin City Council passed an ordinance requiring ride-sharing companies like Uber and Lyft to:

- Conduct fingerprint-based background checks on their drivers (similar to those required for taxi drivers)
- Restrict pick-up and drop-off locations in certain areas
- Collect data and adhere to other safety and operational standards

Uber and Lyft objected to the fingerprint requirement, arguing that their own background checks were sufficient and that fingerprinting was unnecessarily burdensome, potentially deterring drivers from signing up.

In response to the ordinance, Uber and Lyft, along with their supporters, gathered enough signatures to trigger a public referendum (Proposition 1) in May 2016, hoping to overturn the city's regulations. Outcome: Proposition 1 was defeated, with approximately 56% of Austin voters choosing to uphold the city's regulations. Impact: Following the vote, both Uber and Lyft announced that they would suspend operations in Austin, effective May 9, 2016.

However, many residents expressed frustration with the lack of options and higher prices compared to the services provided by Uber and Lyft. The absence of the two giants also created logistical challenges for large-scale events like SXSW and the Austin City Limits music festival.

In May 2017, the Texas Legislature passed House Bill 100, a statewide law that overrode local ridesharing regulations and established uniform rules across Texas. Notably, the new law removed the requirement for fingerprint-based background check and gave regulatory authority over ridesharing companies to the state rather than individual cities. With this new statewide framework, Uber and Lyft resumed operations in Austin soon after the bill was signed into law.

3.2 Data description

This project comprises two data-tables, `platform_economy_reviews.csv` and `platform_economy_restaurants.csv`. The former contains restaurant review tone, while the latter contains restaurant workers' career spells.

3.2.a Reviews

The data-table `platform_economy_reviews.csv` contains 58,227 reviews regarding a sample of 1,107 restaurants in Texas. Below is the codebook for the dataset:

- **id** : Restaurant numeric identifier
- **yelp_tier** : The tier of the restaurant according to Yelp (in "$")
- **time**: Review timestamp in "YYYY-MM" format
- **bad_service**: Percentage of month t Yelp reviews that were negative about the service This measure was achieved by training an aspect-based (i.e., service-based) sentiment classifier on the corpus 58,227 reviews
- **bad_food**: Percentage of month t Yelp reviews that were negative about the food. This measure was achieved by training an aspect-based (i.e., food-based) sentiment classifier on the corpus 58,227 reviews
- **location**: Restaurant location (either Austin or Dallas)

```
df2 = pd.read_csv("data/platform_economy_reviews.csv")
df2.head()
```

	id	yelp_review	time	bad_service	bad_food	location
0	13	$$	2014-05	0.500000	0.300000	austin
1	13	$$	2014-06	0.600000	0.000000	austin
2	13	$$	2014-07	0.375000	0.375000	austin
3	13	$$	2014-08	0.285714	0.142857	austin
4	13	$$	2014-09	0.200000	0.000000	austin

Figure 2 shows the distribution of Yelp reviews regarding restaurants in Austin (which was affected by the Uber and Lyft exit) and Dallas (which was not affected). These data offer ready-to-analyze information regarding the quality of service and food in the restaurants over time.

```
gr = df2.groupby(["time", "location"])
ds = pd.DataFrame(gr.size().reset_index(name="counts"))
fig = plt.figure(figsize=(8, 4))
ax = fig.add_subplot(111)
for _, color in zip(["austin", "dallas"], ["red", "blue"]):
    tmp = ds.loc[ds["location"] == _, :]
    ax.plot(
        tmp.time,
        tmp.counts,
        color=color,
        label=_.title(),
    )
xticklabels = []
for _ in ax.get_xticklabels():
    if "-05" in _.get_text():
        xticklabels.append(_.get_text())
    else:
        xticklabels.append("")
ax.set_ylabel("Counts of Yelp reviews")
ax.set_xticklabels(xticklabels, rotation=90, horizontalalignment="center")
ax.legend(loc="best")
ax.vlines(x="2016-05",    ymin=0,    ymax=ds["counts"].max(),    color="purple",
linestyle="--")
ax.text(x="2016-02",   y=0,   s="Uber   &   Lyft\nexit   Austin",   rotation=90,
color="purple")
ax.vlines(x="2017-05",    ymin=0,    ymax=ds["counts"].max(),    color="green",
linestyle="--")
ax.text(
   x="2017-02", y=0, s="Uber & Lyft\nre-entry Austin", rotation=90, color="green"
)
ax.grid(True, which="major", axis="y", linestyle="--", linewidth=0.5)
plt.show()
```

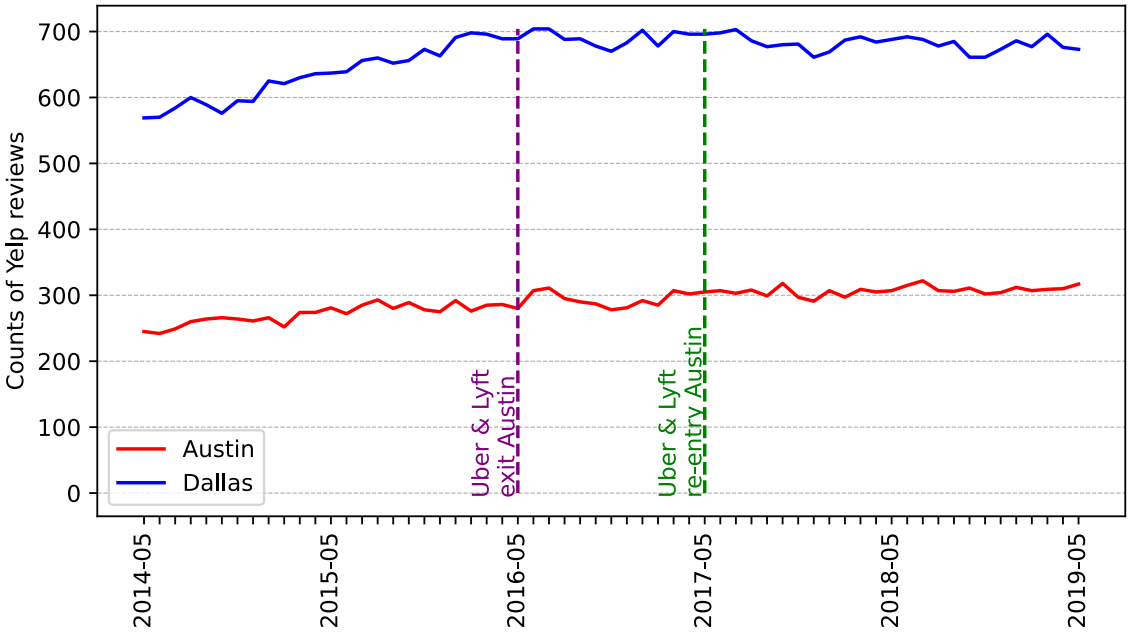

Figure 2: Distribution of Yelp reviews regarding restaurants in Austin and Dallas

3.2.b Labor mobility

A second, related dataset, `platform_economy_mobility.csv`, contains the career spells of 193 restaurant workers. Below is the codebook for the dataset:

- **employee**: Employee numeric identifier
- **time**: Even timestamp in `%YYYY%MM` format
- **tenure**: Worker tenure in the restaurant in months
- **quit**: Logical value whether the employee quit his or her job at the restaurant after the last month indicated by the `time`
- **role**: Worker role in the restaurant
- **avg_hourly_wage**: Worker average hourly wage in that calendar month
- **dma**: Market area of restaurants
- **restaurant**: Restaurant unique identifier (linked against the previous data-table)
- **restaurant_category**: Restaurant category, e.g., "Quick Service," "Fast Casual," "Casual Dining," "Upscale Casual," and "Fine Dining"

```
df2 = pd.read_csv("data/platform_economy_mobility.csv")
df2.head()
```

	em-ployee	time	tenure	quit	avg_ho	urly_wage	dma	restau-rant	restau-rant_-cat
0	1	2017-10	5	False	Prep Cook	14.917600	AUSTIN	3	Upscale Casual
1	1	2017-11	6	False	Prep Cook	15.008746	AUSTIN	3	Upscale Casual
2	1	2017-12	7	False	Prep Cook	14.610148	AUSTIN	3	Upscale Casual
3	1	2017-6	1	False	Line Cook	15.906409	AUSTIN	3	Upscale Casual
4	1	2017-7	2	False	Prep Cook	15.393874	AUSTIN	3	Upscale Casual

3.3 Data visualization problem

Create two (2) plots addressing the following questions: “Does the exit of Uber and Lyft from Austin, Texas, affect the quality of service and food in local restaurants?” And “through which mechanism(s)?”

4 Submission package

The submission package consists of:

- An executive summary containing the two plots and a 600 word, companion text addressing the above-described business problems
- The computer code that allows me to fully reproduce your charts (being, R, Python, Julia, Rust, Stata, C++, Java, etc.). The code should be well-commented and easy to read. Non-reproducible charts will not be graded.