LEUNG WING YAN
18 September 2020

Data Analyst Program

Individual Project

Long Story Short

Uber allows passengers to request the closest available Uber cars via the online platform and watch the car 's progress towards their location instead of waiting in the street or call taxis' services. The transactions are done via the app, so passengers can travel without cash or cards. The company does not own cars, but signs up private drivers. It sets the price of the ride according to the supply and demand. 70-80% of each fare goes to the driver and the rest is kept by Uber.

Course Goals

S7 - Communicate problems, recommendations and insights adapted to the intended target audience

Here are two sets of questions to be answered in this data analysis. My intended target audience is Uber drivers.

1. **What are the most popular pickup and drop off point? (based on district and time)**

- Strategy: How can Uber drivers maximise their profits accordingly? - Recommendation

- Strategy2: Does it have anything to do with weather, holidays, and major event? - Insight

According to Uber's first economics paper , the median Uber driver earned $30.35 per hour in New York, after Uber's take, but before insurance and on-the-job expenses like gas and vehicle maintenance. While Uber is increasing their commission rate from 10% to 15% and now 20% in most of their markets and using the price of the trip as an avenue to beat the competitors, the drivers' income is going down. With all the fixed costs absorbed by drivers, the only way out is to spend more hours in a car. Such business model has been criticised not being sustainable.

Uber has collaborated on research papers with economic superstars like Levitt and former Obama adviser Alan Krueger. The Uber's economics team (Ubernomics) has exaggerated how much the driver can earn, but the driver do not realise how much they are actually spending. So, here comes my second question:

2. **Can Uber drivers earn reasonably by metric (hourly rate)? (Estimated fare is calculated by distance and travel time)**

Conclusion to the above questions

A1:

	PULocationID	Borough	Zone	counts
0	61	Brooklyn	Crown Heights North	855168
1	76	Brooklyn	East New York	687519
2	79	Manhattan	East Village	611765
3	132	Queens	JFK Airport	609634
4	138	Queens	LaGuardia Airport	598407
5	37	Brooklyn	Bushwick South	596288
6	42	Manhattan	Central Harlem North	587589
7	244	Manhattan	Washington Heights South	526249
8	231	Manhattan	TriBeCa/Civic Center	519359
9	161	Manhattan	Midtown Center	504334

Figure 1: The overall most popular pick-up zone

	DOLocationID	Borough	Zone	counts
0	265	Unknown	NaN	1475876
1	61	Brooklyn	Crown Heights North	865096
2	132	Queens	JFK Airport	740448
3	76	Brooklyn	East New York	713482
4	138	Queens	LaGuardia Airport	638658
5	37	Brooklyn	Bushwick South	613069
6	42	Manhattan	Central Harlem North	549653
7	244	Manhattan	Washington Heights South	529838
8	79	Manhattan	East Village	525862
9	89	Brooklyn	Flatbush/Ditmas Park	507319

Figure 2: The overall most popular drop off zone

Human beings are better at processing images rather than text or numbers.

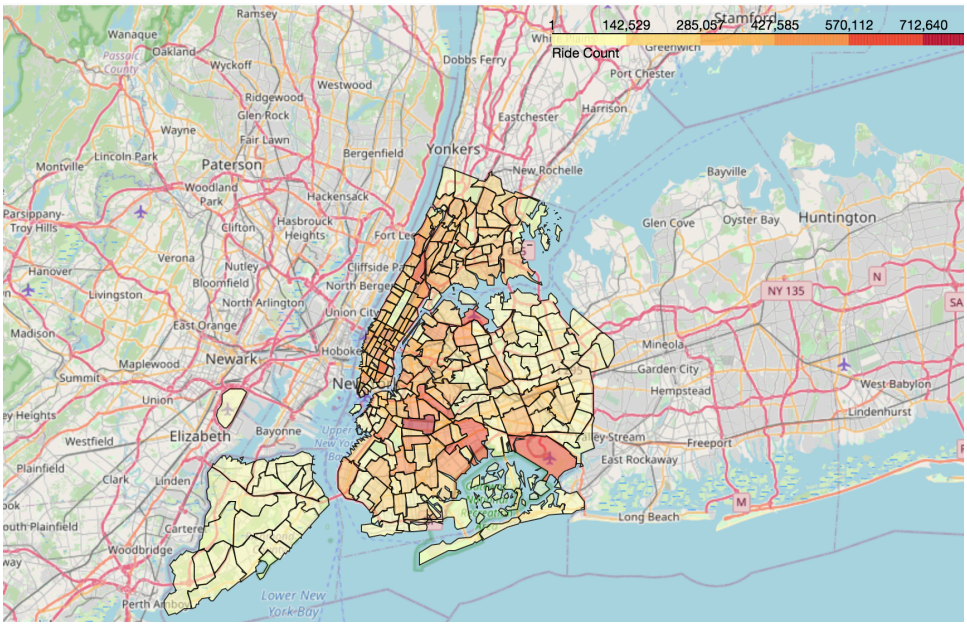

Figure 3: Total pick up count in New York

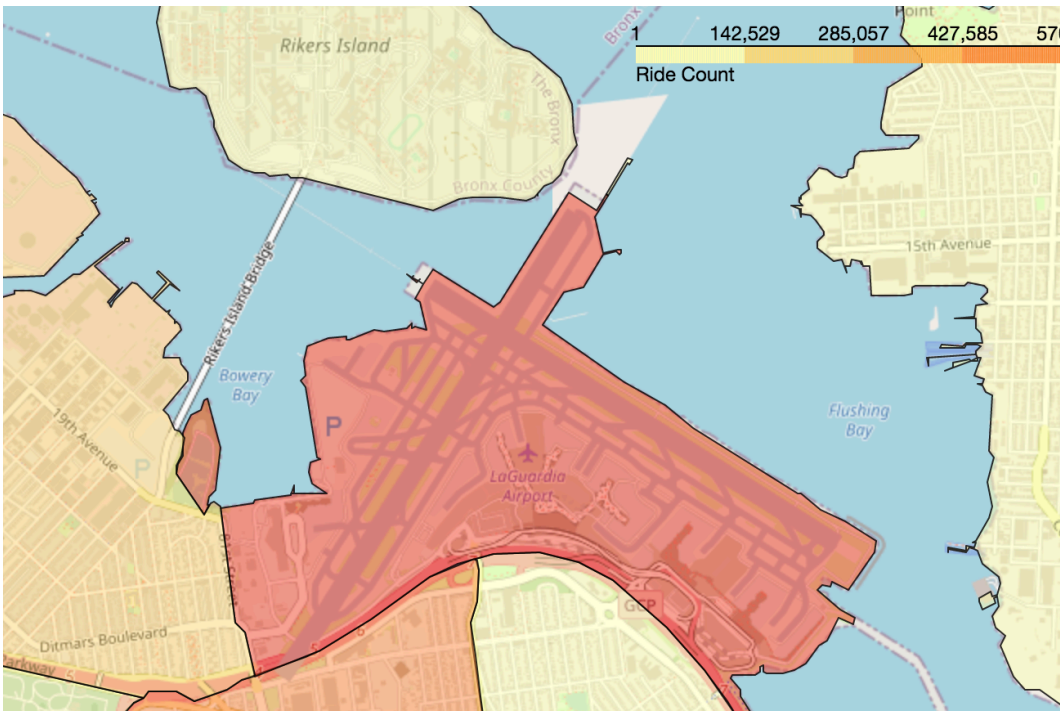

Figure 4: Zoom-in

We can spot out the most popular pick up zone by zooming in. For example, LaGuardia Airport

A/B Testing:

We believe that by placing more Uber cars around the most pickup zone such as Crown Heights North, more rides can be taken by Uber drivers, thus increasing their profits.

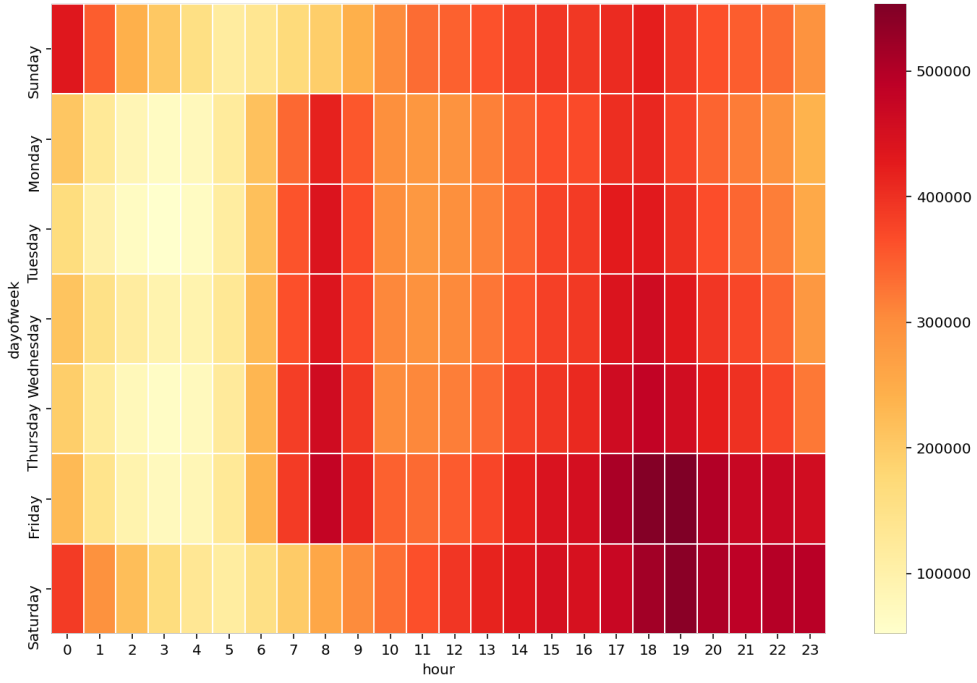

Figure 5: Most pick up by hour by day of week

In Figure 5, If we view the most pick up by time, it is noticeable that most users took the ride in the morning (7-9am) and in the evening (17-19pm) from Monday to Friday. The users probably took Uber to go to work and off from work. On Friday, it still existed a high demand after 7pm (until midnight). The similar pattern for night time was also observable on Saturday. On Sunday, the demand for night time period was less than that on Saturday. It is probably because people had to work on Monday and did not want to be too late to home.

Since the pattern between weekdays and weekends are quite different, let us drill down to the most pick up and most drop off location at day of week level.

	PULocationID	Borough	Zone	counts
0	61	Brooklyn	Crown Heights North	278221
1	79	Manhattan	East Village	246763
2	37	Brooklyn	Bushwick South	214164
3	76	Brooklyn	East New York	204596
4	42	Manhattan	Central Harlem North	185349
5	148	Manhattan	Lower East Side	183667
6	255	Brooklyn	Williamsburg (North Side)	182576
7	132	Queens	JFK Airport	178559
8	7	Queens	Astoria	178400
9	244	Manhattan	Washington Heights South	160538

Figure 6: Most pick up location during weekends

	DOLocationID	Borough	Zone	counts
0	265	Unknown	NaN	418254
1	61	Brooklyn	Crown Heights North	281640
2	37	Brooklyn	Bushwick South	217450
3	79	Manhattan	East Village	211602
4	76	Brooklyn	East New York	211550
5	132	Queens	JFK Airport	209207
6	42	Manhattan	Central Harlem North	174410
7	7	Queens	Astoria	173216
8	225	Brooklyn	Stuyvesant Heights	163232
9	244	Manhattan	Washington Heights South	162756

Figure 7: Most drop off location during weekends

The most rides happened on weekdays and weekends were more or less the same. More rides happened around the **commercial zone** such as TriBeCa/Civic Centre in Manhattan during weekdays while around the **residential, leisure and entertainment area** such as Lower East Side zone during weekends.

A/B Testing 2:

We believe that by placing more Uber cars around the commercial zone during weekdays while placing more around the residential, leisure and entertainment area during weekends, there will be an increasing rides taken by Uber drivers, thus increasing their income.

A2:

Data limitation

Since the dataset does not include Uber drivers or users information, we cannot calculate the hourly rate with the corresponding ride details. Some people had uploaded their own Uber trip online. However, it is quite expensive to collect a reasonable amount of samples online. The alternatives would be calculating the hourly rate based on the estimated fare per ride for UberX and then taking the average value to get a general idea of the hourly rate.

We assume that the idling time of Uber cars should be excluded by this method.

summary	hourly_rate_USD
count	46696376
mean	68.1936969291813
stddev	42.11805065438414
min	42.560776621747166
max	3095.351984110817

Table 1

From Table 1, the average hourly rate earned by Uber drivers was $68.19 per hour and the minimum rate was $42.56 per hour. Since Uber would take 20-30% from the drivers' revenue, the average hourly rate would be $47.73 ($68.19*0.7) and the minimum hourly rate would be $29.79 after the Uber's take. Taking a further step (**Table 2**), 25% of Uber driver earned less than or equal to $38.45 ($54.93*0.7) after the Uber takes, but before insurance and on-the-job expenses.

summary	hourly_rate_USD
25%	54.92534162215543
75%	75.39064631797521

Table 2

According to the State of Working America Wages 2019, the median **wage** in **2019** is $19.33 per hour. If taking insurance and on-the-job expenses like fuel into account, it is quite likely that certain percentage of the Uber drivers did not earn reasonably.

S10 - Plan, execute and identify resources for carrying out experiments to draw data-informed conclusions

Data Collection

The Dataset was downloaded from NYC Open Data [https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page]

Supplementary data

The original dataset only contains PULocationID, DOLocationID, but no correspondent names and coordinates of the places are provided. Two additional datasets regarding NYC Taxi zones and the coordinates of the zone in multi-polygon were imported.

Field Name	Description
Hvfhs_license_num	The TLC license number of the HVFHS base or business As of September 2019, the HVFHS licensees are the following: • HV0002: Juno • HV0003: Uber • HV0004: Via • HV0005: Lyft
Dispatching_base_num	The TLC Base License Number of the base that dispatched the trip
Pickup_datetime	The date and time of the trip pick-up
DropOff_datetime	The date and time of the trip drop-off
PULocationID	TLC Taxi Zone in which the trip began
DOLocationID	TLC Taxi Zone in which the trip ended
SR_Flag	Indicates if the trip was a part of a shared ride chain offered by a High Volume FHV company (e.g. Uber Pool, Lyft Line). For shared trips, the value is **1**. For non-shared rides, this field is **null**.

Data limitation

Compared to Taxi data in which the records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemised fares, rate types, payment types, and driver-reported passenger counts, Uber data provides far less datum. For example, itemised fares, rate types, payment types, and driver-reported passenger counts are not provided.

Moreover, no exact coordinates of pick-up and drop-off points are available in the datasets. The less-than-ideal alternative would be picking the first set of coordinates of the zone so that we can still calculate the trip distance by using Google map API - distance matrix.

Uber charges riders per mile and minute whether they're moving or idling. The estimated fare for UberX can be calculated by the following formula:

```
base_fare = 2.55
per_minute = 0.35
per_mile = 1.75
min_fare = 7

estimated_fare = base_fare + data.duration_min * per_minute + data.distance_km *0.6213* per_mile
```

Take below trip as an example:

```
46 52710,129,82,41449,"(40.7660516539999, -73.87586503899996)","(40.744071717999944, -73.86768504899995)","
{'destination_addresses': ['370006 Junction Blvd, Flushing, NY 11368, USA'], 'origin_addresses': ['94 St/24 Av, Queens, NY
11369, USA'], 'rows': [{'elements': [{'distance': {'text': '2.6 km', 'value': 2562}, 'duration': {'text': '11 mins', 'value':
682}, 'status': 'OK'}]}], 'status': 'OK'}"
```

Figure 8: Trip distance and trip duration extracted from google map API

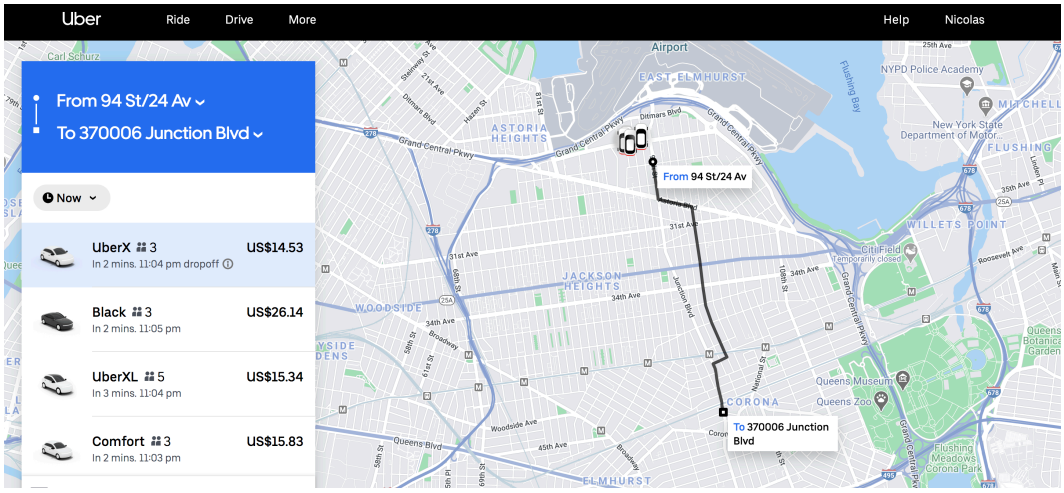

Figure 9: The fares for different types of Uber cars from point to point

As you can see, the accuracy is very low. The actual fare is usually higher than the calculated one. The reason is that Uber adopts surge pricing model, which imposes higher fares during the period of higher rider demand such as peak hour and bad weather. The best accuracy would be using Uber API to retrieve the estimated fare. However, we would use the above formula to get the estimated fare for simplicity.

C1 - Process and analyze data in such a way that it leads to further learning and professional development

- Strategy2: Does it have anything to do with weather, holidays, and event? - Insight

Apart from the peak hour, I would like to investigate if there are other factors such as weather, holidays and major events which can affect the Uber fare.

In figure 10, *the average number of Uber rides was 441,347 . It is likely that the major holidays such as New Year Day boosted the number of ride trip (530, 257 in total) on 1 January 2020. Heavy rainy day also increased the total rides on 13 February 2020. The peak of total rides per day happened on 29 February 2020. The outbreak of coronavirus started in early March and the number of total rides started to fall dramatically. It fell to the lowest point (92,514 total rides) where the coronavirus was at the peak and then went up gradually.

* The average was taken by the data between 01-01-2020 - 01-03-2020 (excluding the coronavirus effect)

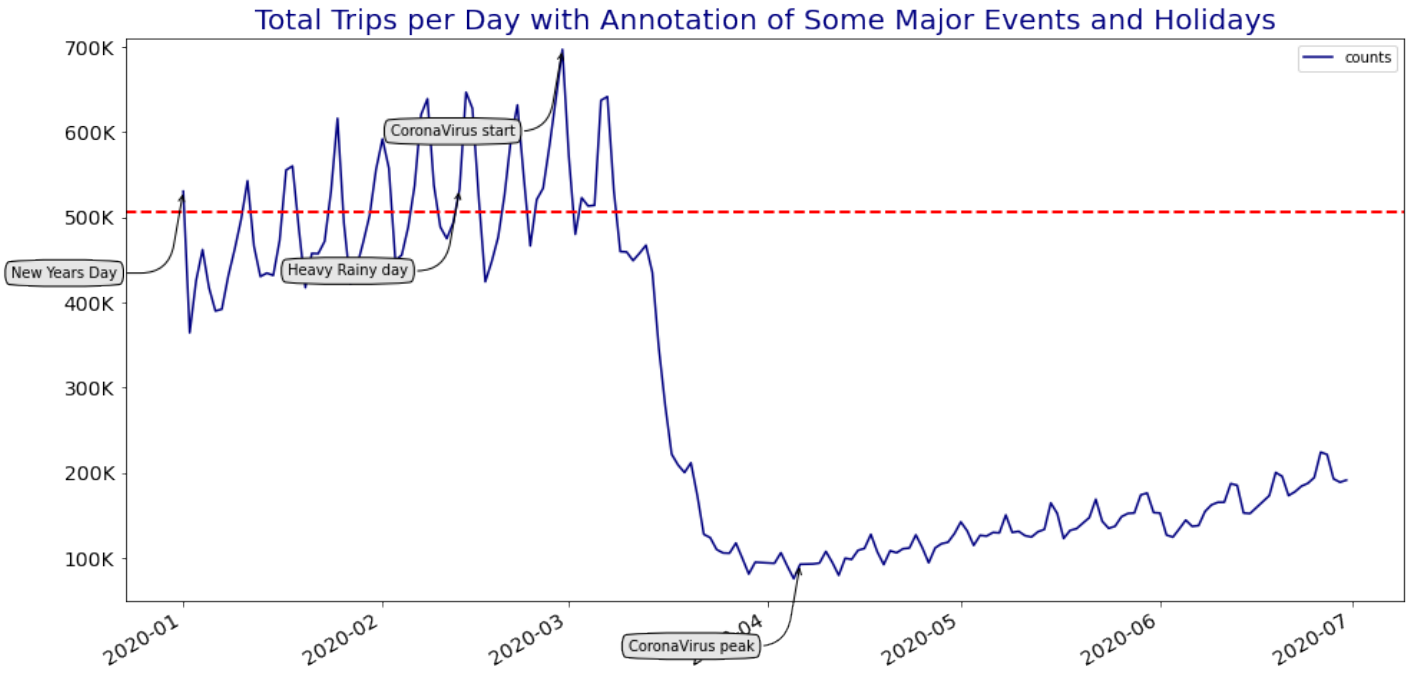

Figure 10: Number of ride trips with major events and holidays

C3 - Independently be able to work methodically and flexibly in various projects and processes

```
from pyspark.sql import SQLContext

sqlContext = SQLContext(sc)

parquetFile = sqlContext.read.parquet('fhvhv_tripdata_2020-01_to_2020-06.parquet.gzip')

#parquetFile.columns
```

```
from pyspark.sql.functions import to_timestamp, to_date
from pyspark.sql.functions import dayofmonth, dayofweek, hour, minute

pf = parquetFile.select(
    to_timestamp(parquetFile.pickup_datetime, 'yyyy-MM-dd HH:mm:ss').alias('pickup'),
    to_timestamp(parquetFile.dropoff_datetime, 'yyyy-MM-dd HH:mm:ss').alias('dropoff'),
    (to_timestamp(parquetFile.dropoff_datetime, 'yyyy-MM-dd HH:mm:ss').cast('long') \
     - to_timestamp(parquetFile.pickup_datetime, 'yyyy-MM-dd HH:mm:ss').cast('long')).alias('travel_time_seconds'),
    parquetFile.DOLocationID,
    parquetFile.PULocationID,
    parquetFile.SR_Flag,
    to_date(parquetFile.pickup_datetime).alias('pickup_date')
)

pf = pf.withColumn('dayofweek', dayofweek(pf.pickup_date))
pf = pf.withColumn('dayofmonth', dayofmonth(pf.pickup_date))
pf = pf.withColumn('hour', hour(pf.pickup))
pf = pf.withColumn('minute', minute(pf.pickup))
```

noun: 拾音器, 拾音, 拾起, 起色, 唱頭
adjective: 拾, 隨便

Figure 11: Using PySpark to load and proceed the data

The Uber dataset is quite big with around 53 million rows (data period: 01-01-2020 to 30-06-2020). It took very long time and memory to load the csv file and proceed it further. To resolve the performance issue, I first converted it into parquet file and then use PySpark to proceed it.

My Learning

In this project, I have learnt how to proceed the large amount of data using pyspark. I have solidated my coding skills in python as well as data visualisation skills using matplotlib and seaborn. Now, I can remember the syntax without google it. I have brushed up my debugging skills. Moreover, I have learnt how to visualise the geospatial data by using libraries such as geopandas and folium which is nice to be put in my toolbelt. I am able to conduct data analysis based on limited data available and find a way (e.g. using Google map API) to supplement my data. My critical analysis skills is also enhanced by understanding the data better and further proceeding it in such a way that leads to finding a better insight. I believe that now, I have equipped with good amount of data analytical skills compared to one year ago.

My Contact

E-mail: miki.lwy@gmail.com

LinkedIn: https://www.linkedin.com/in/wing-yan-leung-63b02056/

Github: https://github.com/miki-lwy

Reference

The Guardian Australia. Sam Levin. Uber drivers often make below minimum wage, report finds. Available at: https://www.pressreader.com/australia/the-guardian-australia/20180306/282093457244589 [07/03/2018]

TLC Trip Record Data. Available at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page [14/09/2020]

Data Dictionary – High Volume FHV Trip Records. Available at: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_hvfhs.pdf [19/09/2019]

Economic Policy Institute. State of Working America Wages 2019. Elise Gould • February 20, 2020. Available at: https://www.epi.org/publication/swa-wages-2019/ [20/02/2020]

NYC Open Data. COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths. Available at: https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3 [16/09/2020]

Weather Data Services. New York. Available at: https://www.visualcrossing.com/weather/weather-data-services?pln=plan_GqkbrxoxU5fY8e#/loadingscreen [30/08/2020]