Uber Data

A project by: The A Team

Your drivers for today’s journey will be: Marta, Brad, Kareem, Mariama and Maliha.

By the time you reach your destination you should know …

Project Outline

➔ *Why Uber*

- Uber has become the most popular choice for travel across the world.
- In 2021, there were 118 million users in over 80 countries.

- Exploration of Uber data in New York between 2009 and 2014.
- Exploration of the consumer behaviour.

Questions we asked:

1. Does the time/date affect how often people order Uber's?
2. How has Uber prices changed over time?
3. What distance are people using Uber's for?
4. What are the most common amount of passengers for an Uber

Data Exploration

- **Uber Data – csv file**
- **Uber trips in New York 2009 - 2015**
- **Kaggle Dataset -** Uber Fares Dataset | Kaggle

```
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   id                 200000 non-null  int64
 1   key                200000 non-null  object
 2   fare_amount        200000 non-null  float64
 3   pickup_datetime    200000 non-null  object
 4   pickup_longitude   200000 non-null  float64
 5   pickup_latitude    200000 non-null  float64
 6   dropoff_longitude  199999 non-null  float64
 7   dropoff_latitude   199999 non-null  float64
 8   passenger_count    200000 non-null  int64
dtypes: float64(5), int64(2), object(2)
memory usage: 13.7+ MB
```

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5

Clean-up process

UBER

```
# Delete column "Key"
del uber_file_df["key"]

# Delete rows  with 0 longitude and longitude value
df2 = uber_file_df[ (uber_file_df['pickup_longitude'] == 0) & (uber_file_df['pickup_latitude'] == 0)
                  & (uber_file_df['dropoff_longitude'] == 0)
                  & (uber_file_df['dropoff_latitude'] == 0)].index
uber_file_df.drop(df2 , inplace=True)

# Drop N/A values in the dataset
uber_file_df = uber_file_df.dropna(how='any')

# Delete rows  with 0 passengers and fare amount value
uber_file_df.drop(uber_file_df[uber_file_df['passenger_count'] == 0].index, inplace = True)
uber_file_df.drop(uber_file_df[uber_file_df['passenger_count'] == 208].index, inplace = True)
uber_file_df.drop(uber_file_df[uber_file_df['fare_amount'] < 1].index, inplace = True)

uber_file_df.head(15)
```

```
# Removing rows
uber_file_df.drop(uber_file_df[uber_file_df['Distance'] > 100].index, inplace = True)
uber_file_df.drop(uber_file_df[uber_file_df['Distance'] == 0].index, inplace = True)
```

```
# Delete coordinates as we have now calculated distance
del uber_file_df["pickup_latitude"]
del uber_file_df["pickup_longitude"]
del uber_file_df["dropoff_longitude"]
del uber_file_df["dropoff_latitude"]
```

```
# Creating new columns Month Date Day Hour Day of week
uber_file_df['pickup_datetime'] = pd.to_datetime(uber_file_df['pickup_datetime'])

uber_file_df['Year'] = uber_file_df['pickup_datetime'].apply(lambda time: time.year)
uber_file_df['Day'] = uber_file_df['pickup_datetime'].apply(lambda time: time.day)
uber_file_df['Hour'] = uber_file_df['pickup_datetime'].apply(lambda time: time.hour)
uber_file_df['Month'] = uber_file_df['pickup_datetime'].apply(lambda time: time.month)
uber_file_df['Day of Week'] = uber_file_df['pickup_datetime'].apply(lambda time: time.dayofweek)
uber_file_df['Day of Week_number'] = uber_file_df['pickup_datetime'].apply(lambda time: time.dayofweek)
uber_file_df['counter'] = 1

days = {0:'Mon',1:'Tue',2:'Wed',3:'Thu',4:'Fri',5:'Sat',6:'Sun'}
uber_file_df['Day of Week'] = uber_file_df['Day of Week'].map(days)
```

Final Data Frame

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5
5	44470845	2011-02-12 02:27:09.0000006	4.9	2011-02-12 02:27:09 UTC	-73.969019	40.755910	-73.969019	40.755910	1
6	48725865	2014-10-12 07:04:00.0000002	24.5	2014-10-12 07:04:00 UTC	-73.961447	40.693965	-73.871195	40.774297	5
7	44195482	2012-12-11 13:52:00.00000029	2.5	2012-12-11 13:52:00 UTC	0.000000	0.000000	0.000000	0.000000	1
8	15822268	2012-02-17 09:32:00.00000043	9.7	2012-02-17 09:32:00 UTC	-73.975187	40.745767	-74.002720	40.743537	1
9	50611056	2012-03-29 19:06:00.000000273	12.5	2012-03-29 19:06:00 UTC	-74.001065	40.741787	-73.963040	40.775012	1

1. **Raw data from csv file converted to a data frame**

```
# Checking the shape of data
uber_file_df.shape
```

```
(200000, 9)
```

	id	fare_amount	pickup_datetime	passenger_count	Year	Date	Hour	Month	Day of Week	Day of Week_number	counter	Distance
1	27835199	7.7	2009-07-17 20:04:56+00:00	1	2009	17	20	7	Fri	4	1	2.5
2	44984355	12.9	2009-08-24 21:45:00+00:00	1	2009	24	21	8	Mon	0	1	5.0
3	25894730	5.3	2009-06-26 08:22:21+00:00	3	2009	26	8	6	Fri	4	1	1.7
4	17610152	16.0	2014-08-28 17:47:00+00:00	5	2014	28	17	8	Thu	3	1	4.5
5	44470845	4.9	2011-02-12 02:27:09+00:00	1	2011	12	2	2	Sat	5	1	0.0
6	48725865	24.5	2014-10-12 07:04:00+00:00	5	2014	12	7	10	Sun	6	1	11.7
8	15822268	9.7	2012-02-17 09:32:00+00:00	1	2012	17	9	2	Fri	4	1	2.3
9	50611056	12.5	2012-03-29 19:06:00+00:00	1	2012	29	19	3	Thu	3	1	4.9
12	31892535	3.3	2011-05-17 14:03:00+00:00	5	2011	17	14	5	Tue	1	1	0.3
13	13012786	10.9	2011-06-25 11:19:00+00:00	1	2011	25	11	6	Sat	5	1	3.6

2. **Formatted and filtered data frame without null values.**

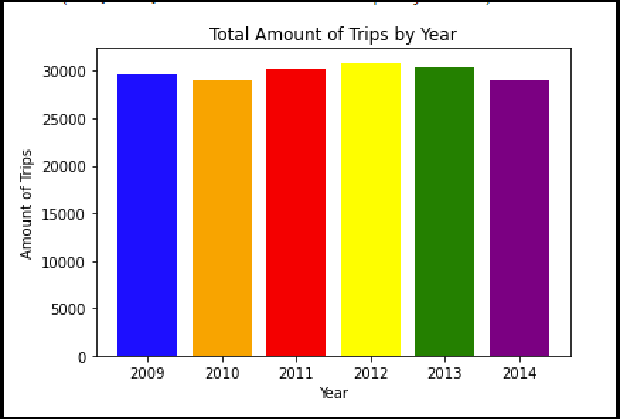
```
# Checking the shape of data after changes
uber_file_df.shape
```

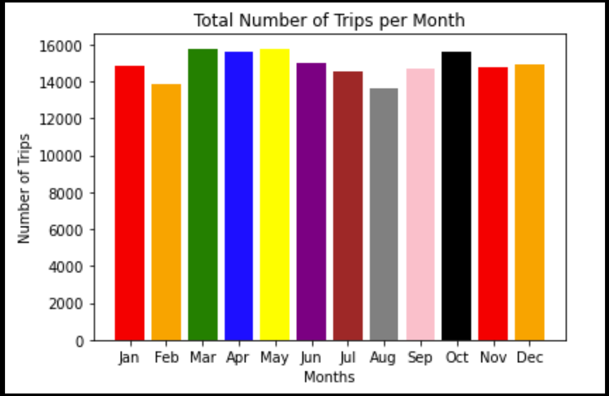
```
(178914, 12)
```

Analysis Process

Time /Date
Does the date affect how often people order Ubers?

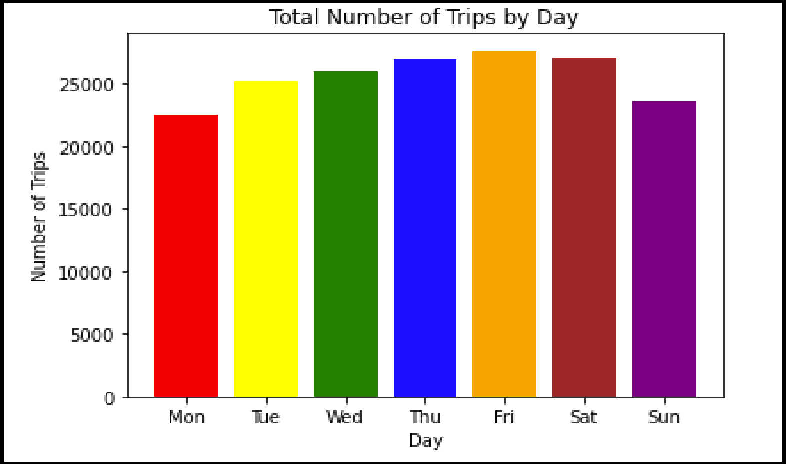
- Initial Assumptions
- Uber Usage would differ over the months
- Slight drop in the months of August and February
- Doesn't show big change in total trips over years





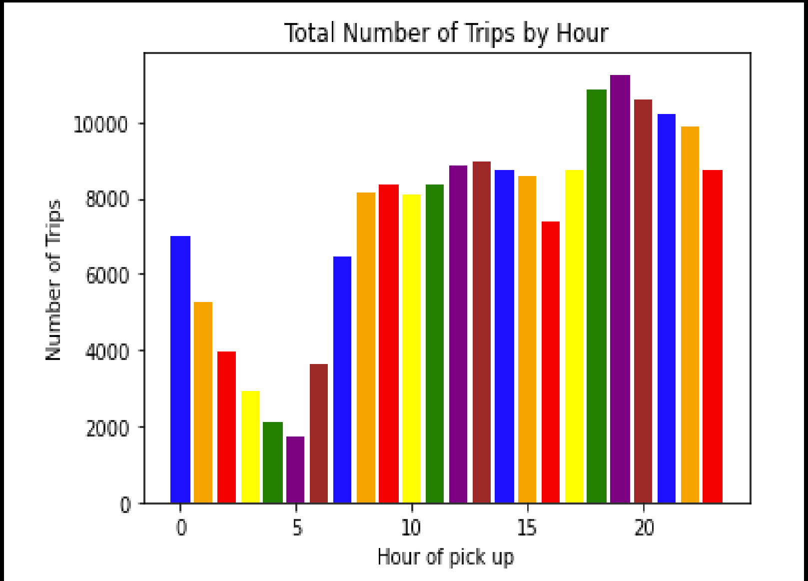
Time /Date
Does the day of the week affect how often people order Ubers?

- We also looked into the days of the week and our initial hypothesis was that Friday and Saturday nights would be when Ubers we’re going to be booked the most because most people don’t work weekend and do fun activities on those day.

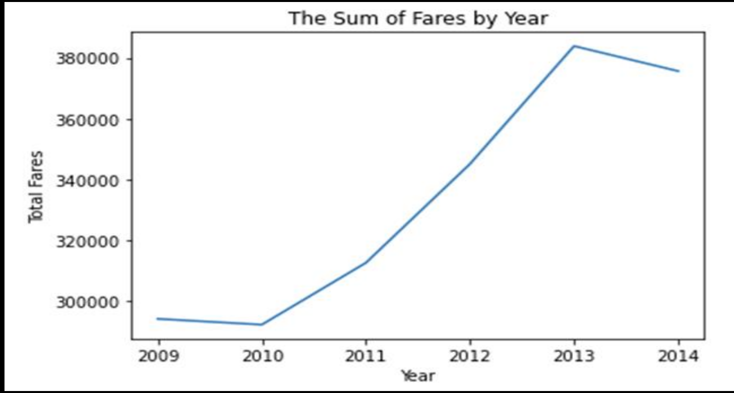


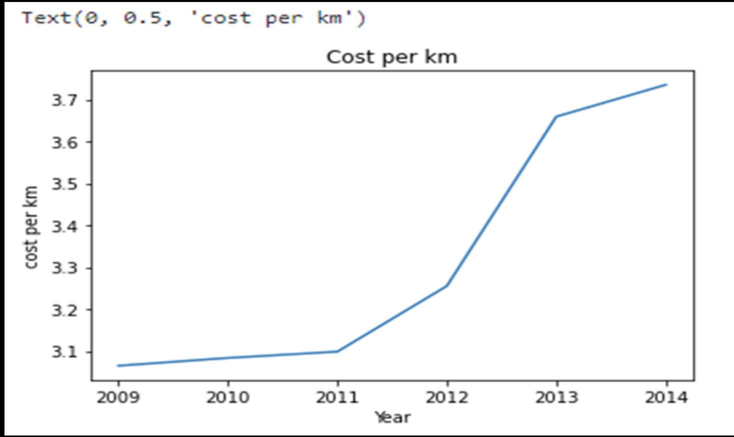
Time /Date
Does the time affect how often people order Ubers?

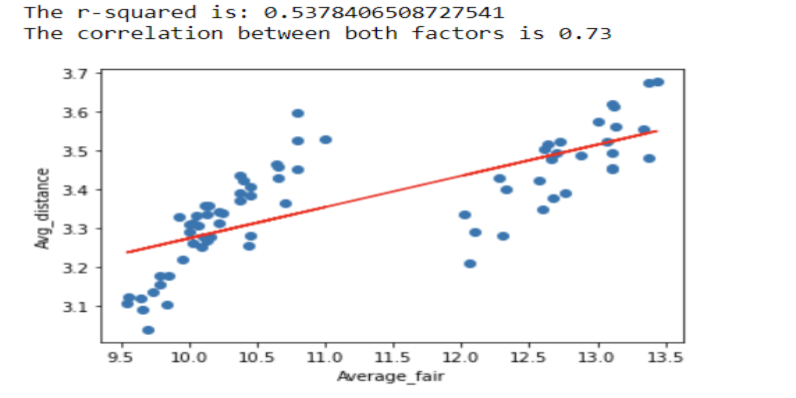
- Evening time in New York
- Peak times between 16:00 – 22:00



Cost

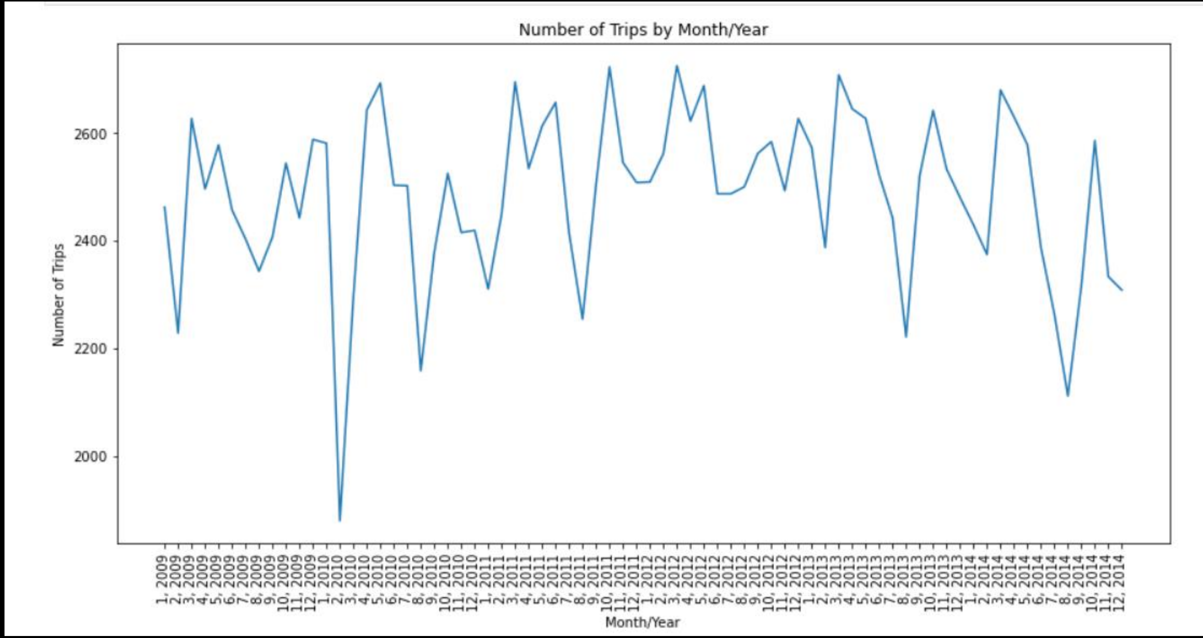






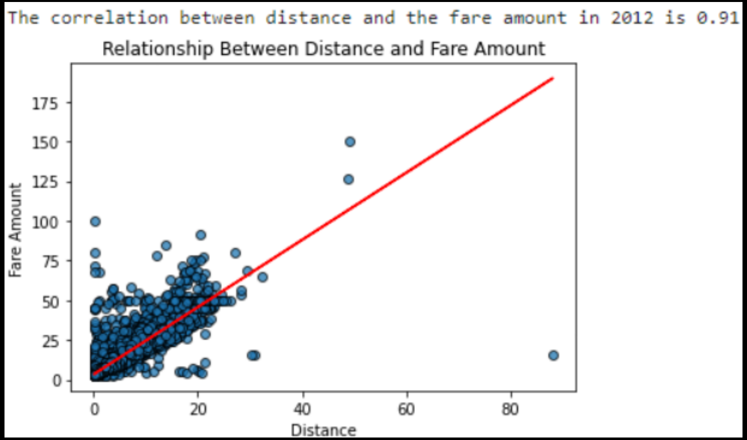
- There is a strong positive correlation between average distance and average fair amount. This make sense as the distance of trip increase, the price of the trip increase. This suggests price can act as a key decision making point while booking an uber.

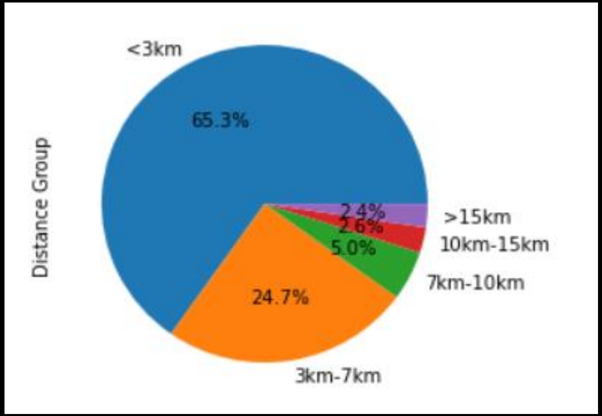
Extreme Weather Event



Distance

- Our data shows that people mainly use Uber for short trips.
- Over 65% of the trips we analyzed were less than 3km in distance
- Nearly 25% were between 3km and 7km in distance
- There’s a very strong positive correlation between the distance and fare amount

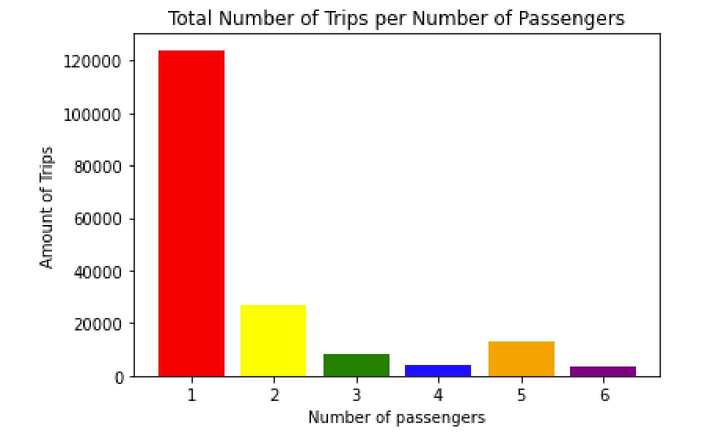


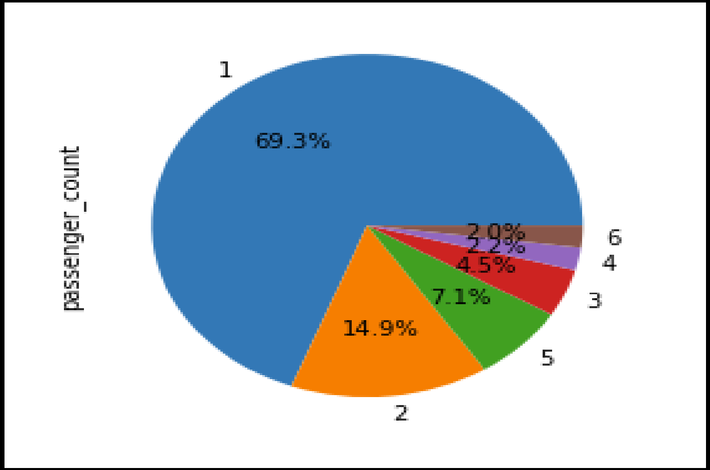


<3km	116771
3km-7km	44268
7km-10km	8986
10km-15km	4627
>15km	4262

Number of Passengers

- Our data shows that people often take uber trips on their own





Conclusions and Implications

- Relationships we looked at
  - Trend of
    - total number of trips/rides
    - total fares collected
    - fare prices adjusted for distance
  - Distribution of
    - total trips per month over 5 years
    - average passenger numbers
    - average distance travelled
  - Time (seasonality) variables (hour of day; day of week; month of year) with
    - total trips
    - average number of passengers
  - Fare price with distance travelled

Conclusions and Implications

- What does this all mean/show?:
  - Company perspective (are we making a profit?):
    - Trip numbers are stable (stagnant?), and not increasing
    - But fares collected are increasing
    - Though not because of increased trips nor increased distances covered
    - Needs company cost data (fuel, repairs, staff, rent etc), and macroeconomic data (inflation) to assess whether this continued growth in fare collected is a profit and whether it is above the levels of inflation
  - Customer perspective (is it worth it?)
    - Fare costs are increasing
    - Numbers using the product is stagnant
- Markets for the company to explore and grow
  - Increase trips
  - Increase distance travelled
  - Increase those with more than one passenger

Thank you