1. Objective of the Analysis

- Perform an EDA exploratory data analysis on the aggregated For-Hire Vehicle FHV data.
- Analyse Uber to other FHV pickups over time
- Summary the insights and recommend the business to determine the next steps for the FHV market

2. Load necessary libraries and packages

In [1]:

```
## Libraries for handling and data manupliation
import numpy as np
import pandas as pd

## Load libraries to read data
import requests as rq


## Libraries for visualization
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

## holiday canedar
from pandas.tseries.holiday import USFederalHolidayCalendar as calendar

## data preprocessing
from sklearn.preprocessing import StandardScaler

## Notebook display settings
pd.set_option("display.precision", 2)
pd.reset_option('display.float_format')
```

3. Load the dataset

In [2]:

```
# URL on the Github where the csv files are stored
github_url = 'https://github.com/fivethirtyeight/uber-tlc-foil-response/blob/master/Aggr

## Reading the FHV aggregated data file in pandas to view in tabular form
fhv_agg_raw = pd.read_excel(github_url, sheet_name='Trips Per Day')

fhv_agg_raw.set_index(['Date'],
                      inplace=True)

fhv_agg_raw.sort_values(by = 'Date',
                        ascending = True,
                        inplace = True)
```

3. Data quaity check

In [3]:

```
## Look at first few rows
```

```
fhv_agg_raw.head(2)
```

Out[3]:

	American	Carmel	Dial 7	Diplo	Firstclass	Highclass	Prestige	Skyline	Lyft	Uber	Yellow Taxis	Gre Ta
Date												
2014-07-01	921	2871	2233	1046	1744	1368	3345	1668	0	21228	440655	381
2014-07-02	1028	2965	2409	1275	2228	1661	3533	1691	0	26480	434416	424

In [4]:
```
## Number of rows and columns
fhv_agg_raw.shape
```

Out[4]:
```
(92, 12)
```

In [5]:
```
## Check if any dates are missing or duplicate
len(fhv_agg_raw.index), fhv_agg_raw.index.nunique()
```

Out[5]:
```
(92, 92)
```

In [6]:
```
## Check the datatype of each column
fhv_agg_raw.dtypes.to_frame(name= 'Datatype')
```

Out[6]:

	Datatype
American	int64
Carmel	int64
Dial 7	int64
Diplo	int64
Firstclass	int64
Highclass	int64
Prestige	int64
Skyline	int64
Lyft	int64
Uber	int64
Yellow Taxis	int64
Green Taxis	int64

In [7]:
```
## Number of missing rows in each column
fhv_agg_raw.isnull().sum().to_frame(name='Number of missing rows')
```

Out[7]:

	Number of missing rows
American	0
Carmel	0
Dial 7	0
Diplo	0
Firstclass	0
Highclass	0

Prestige	0
Skyline	0
Lyft	0
Uber	0
Yellow Taxis	0
Green Taxis	0

Data Quality:

1. The data sheet consists of 3 monthly daily pickups data for various vehicles in NYC.
2. The dataset includes 92 rows of daily pickup counts for 12 companies, including Uber.
3. There are no missing values, and all columns have correct data types. The data range from July to the End of September 2014.

4. Summarzing the data

In [8]:

```
## Data start and end date
fhv_agg_raw.index[0],fhv_agg_raw.index[-1]
```

Out[8]:

```
(Timestamp('2014-07-01 00:00:00'), Timestamp('2014-09-30 00:00:00'))
```

In [9]:

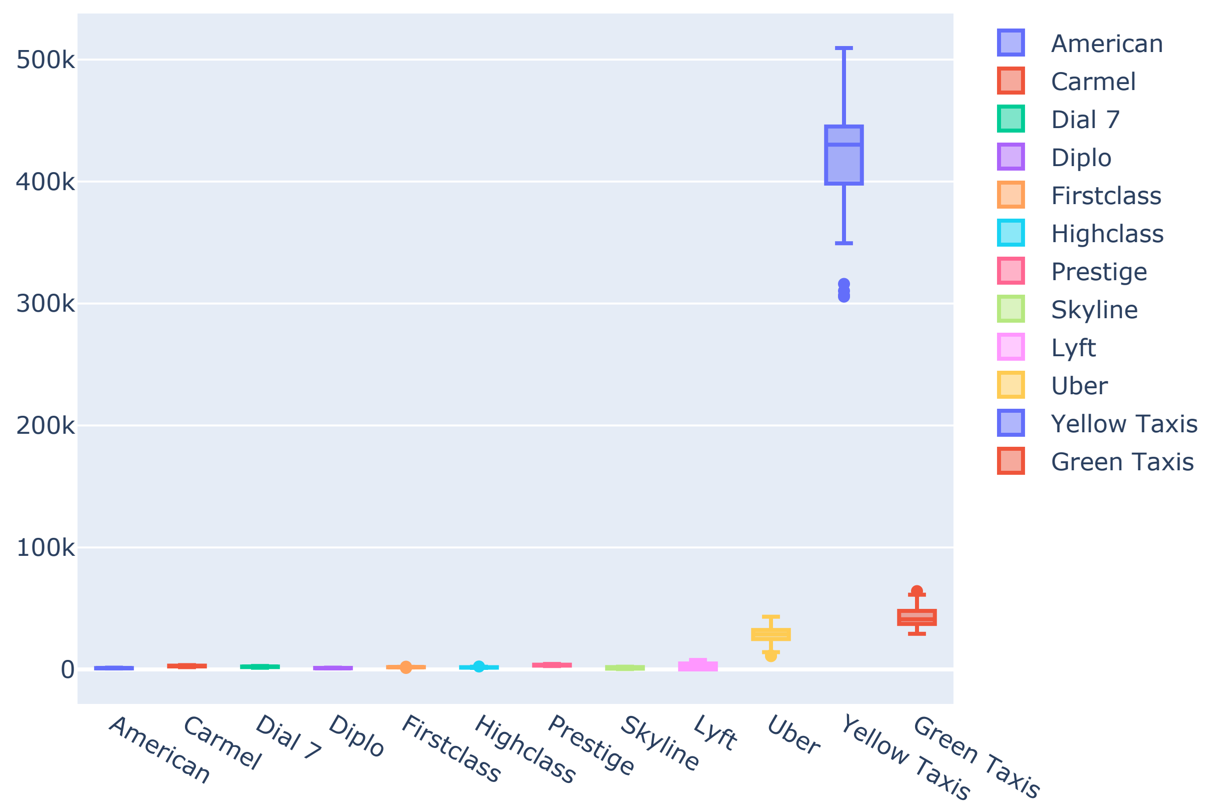
```
## Summary Statistics of the dataset
fhv_agg_raw.describe().T
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
American	92.0	996.87	164.84	768.0	860.00	944.0	1114.50	1440.0
Carmel	92.0	2788.25	382.77	1846.0	2453.00	2882.5	3079.75	3507.0
Dial 7	92.0	2119.48	298.37	1371.0	1912.25	2193.0	2348.25	2795.0
Diplo	92.0	1071.20	163.54	810.0	936.50	1030.0	1227.00	1440.0
Firstclass	92.0	1812.71	147.32	1211.0	1742.00	1802.0	1900.75	2228.0
Highclass	92.0	1651.36	246.79	1315.0	1456.50	1602.5	1816.50	2375.0
Prestige	92.0	3485.23	435.35	2781.0	3111.50	3350.0	3878.25	4470.0
Skyline	92.0	1388.00	629.76	276.0	621.00	1634.5	1897.75	2230.0
Lyft	92.0	2909.79	2443.94	0.0	0.00	2512.5	4876.50	7740.0
Uber	92.0	28842.74	6353.07	10890.0	24922.50	28791.5	32316.25	43205.0
Yellow Taxis	92.0	421398.93	40868.29	305653.0	398623.25	430251.5	444734.75	509480.0
Green Taxis	92.0	43213.74	7500.26	29186.0	37335.75	41118.5	47802.50	64184.0

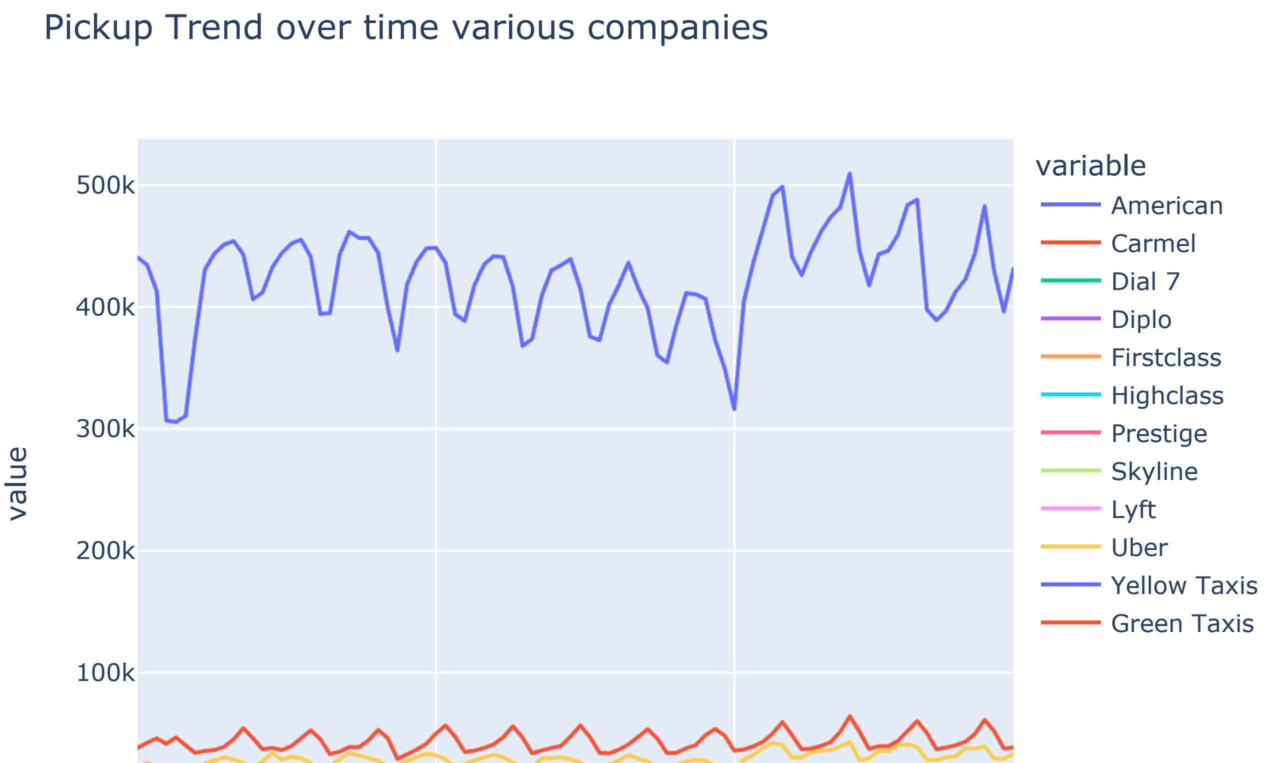
In [10]:

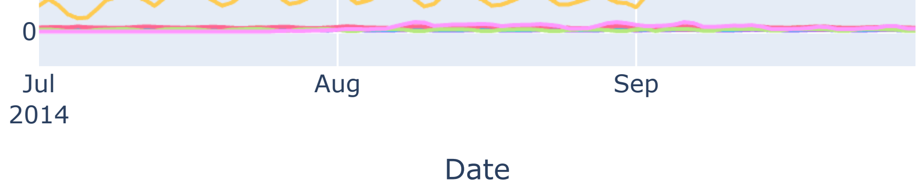
```
## Distribution of pickups for various companies
fig = go.Figure()
for col in fhv_agg_raw:
  fig.add_trace(go.Box(y=fhv_agg_raw[col].values, name=fhv_agg_raw[col].name))

fig.show()
```



In [11]:

```
fhv_agg_raw.reset_index(inplace=True)
fig = px.line(fhv_agg_raw, x="Date", y=fhv_agg_raw.columns,
              hover_data={"Date": "|%B %d, %Y"},
              title='Pickup Trend over time various companies ')
fig.update_xaxes( dtick="M1", tickformat="%b\n%Y")
fig.show()
```





Summary Statistics Info:

1. Based on the summary statistics, Yellow Taxis has the highest count of pickups. This can be attributed to the fact that Yellow Taxis are the only vehicles allowed to pick up passengers anywhere in the city.
2. On the other hand, there are days when Lyft doesn't have any pickups. This could be when this vehicle was not introduced in the market yet.
3. Uber has the 3rd largest market share after Green and Yellow Taxis.
4. It makes sense to exclude the Yellow Taxis hereon as they have XX times pickup rates than the other taxis combined.

4. Feature Engineering and Data Reshaping

In [12]:
```
## excluding Green Taxis from the dataset
df = fhv_agg_raw.loc[:, fhv_agg_raw.columns != 'Yellow Taxis']
```

In [13]:
```
df.head(2)
```

Out[13]:

	Date	American	Carmel	Dial 7	Diplo	Firstclass	Highclass	Prestige	Skyline	Lyft	Uber	Green Taxis
0	2014-07-01	921	2871	2233	1046	1744	1368	3345	1668	0	21228	38167
1	2014-07-02	1028	2965	2409	1275	2228	1661	3533	1691	0	26480	42472

In [14]:
```
## Reshaping the dataframe to long format for the purpose of EDA
df_stack = df.set_index('Date').columns.to_list()
df_stack = df.set_index('Date').stack().reset_index()
df_stack.columns = ['Date', 'vehicle_companies', 'daily_pickups']
```

In [15]:
```
df_stack.head(2)
```

Out[15]:

	Date	vehicle_companies	daily_pickups
0	2014-07-01	American	921
1	2014-07-01	Carmel	2871

In [16]:
```
# ## Calculate market share for all vehicles
df_stack['total_pickups'] = df_stack.groupby(['Date'])['daily_pickups'].transform(sum)
df_stack['daily_pickup_share'] = (df_stack['daily_pickups']/df_stack['total_pickups'])*1
```

In [17]:
```
df_stack.head(2)
```

Out[17]:

	Date	vehicle_companies	daily_pickups	total_pickups	daily_pickup_share

0	2014-07-01	American	921	74591	1.23
1	2014-07-01	Carmel	2871	74591	3.85

In [18]:

```
## Extract date derived features to understand the vehicle perofrmance at weekly and mon
df_stack['week_number'] = df_stack['Date'].dt.isocalendar().week
df_stack['dayofweek'] = df_stack['Date'].dt.dayofweek
df_stack['month'] = df_stack['Date'].dt.month
df_stack['day_name'] = df_stack['Date'].dt.day_name()
df_stack['month_name'] = df_stack['Date'].dt.month_name()
df_stack["is_weekend"] = df_stack.dayofweek > 4
```

In [19]:

```
## Use US public holiday calendar to create a holiday flag
cal = calendar()
holidays = cal.holidays(start=df_stack.Date.min(), end=df_stack.Date.max())

df_stack['Federal_Holiday'] = df_stack['Date'].isin(holidays)
```

In [20]:

```
df_stack.head(2)
```

Out[20]:

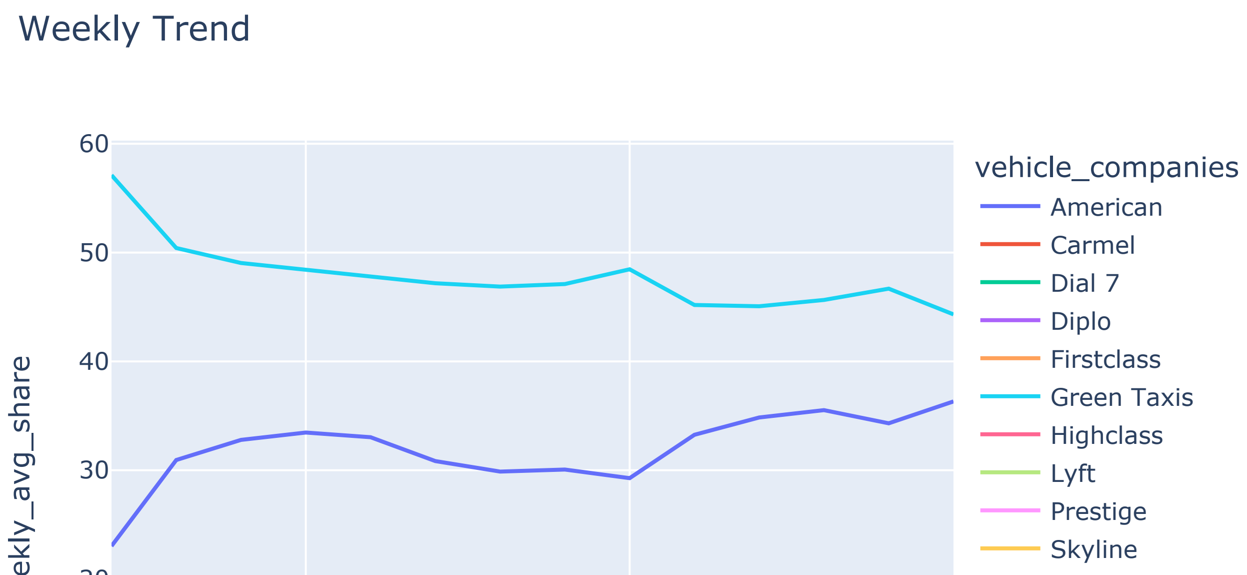
	Date	vehicle_companies	daily_pickups	total_pickups	daily_pickup_share	week_number	dayofweek	n
0	2014-07-01	American	921	74591	1.23	27	1	
1	2014-07-01	Carmel	2871	74591	3.85	27	1	

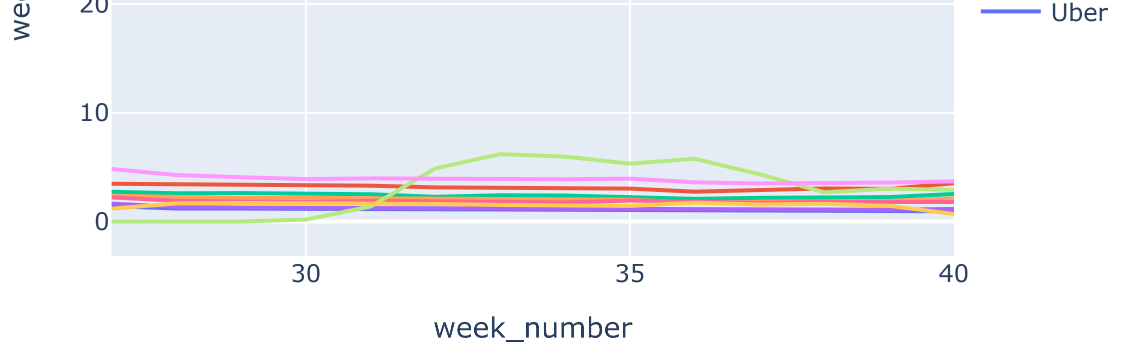
5. Exploratory data analysis

Part 1 : Analysing Uber

In [21]:

```
## calculate weekly avg
weekly_avg = df_stack.groupby(['vehicle_companies','week_number'])['daily_pickup_share']
    columns={'daily_pickup_share' : 'weekly_avg_share'})
fig = px.line(weekly_avg, x="week_number", y="weekly_avg_share",
             color="vehicle_companies",
             title="Weekly Trend")
fig.show()
```

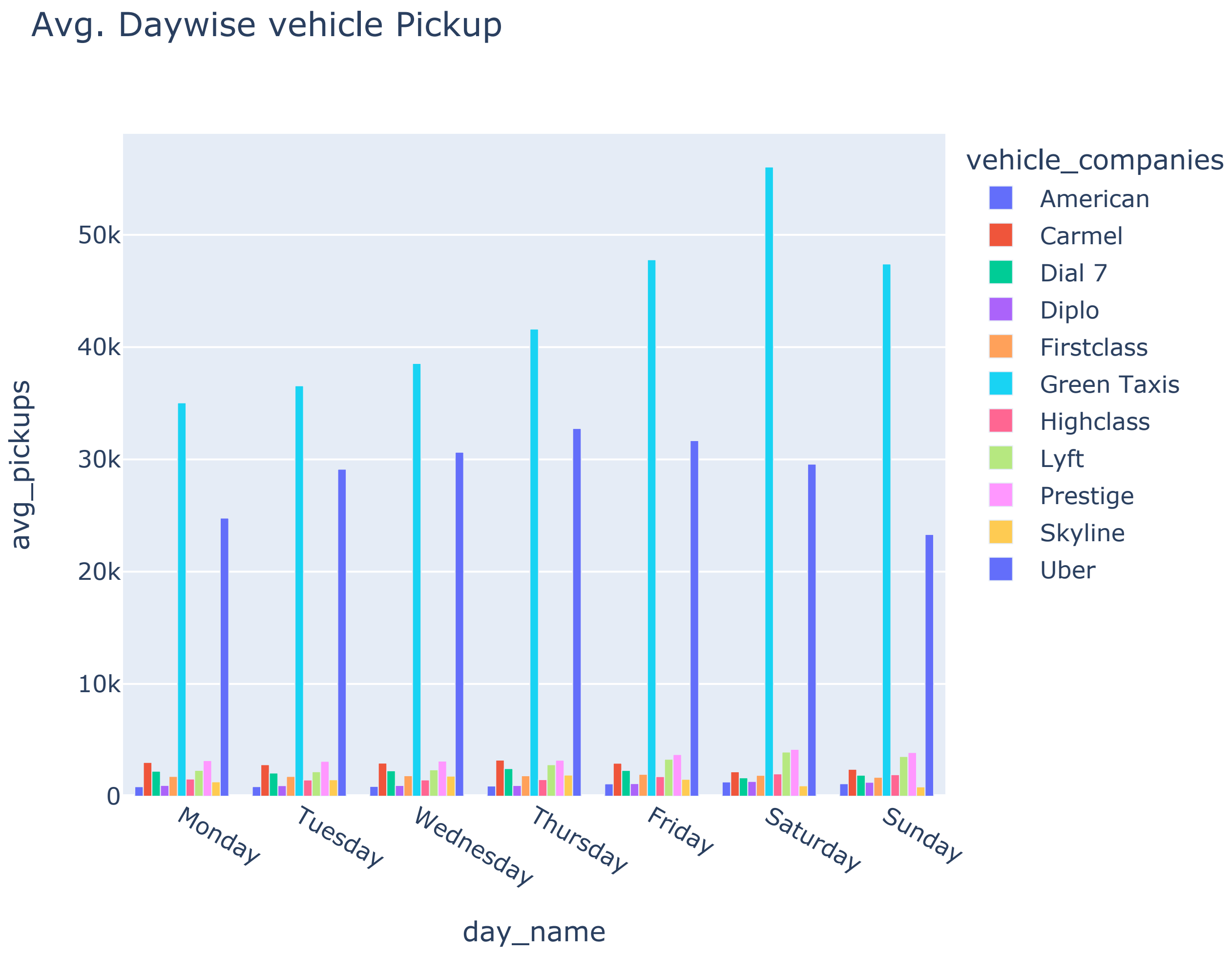




1. Green taxis and Uber continue to dominate the market over different weeks.
2. Over the weeks, the market share gap between green taxis and Uber is becoming smaller and smaller. The graph suggests that Uber is slowly becoming popular.
3. Lyft, although being a late entrant to this market, is also catching up even though it has a < 10% market share.

In [22]:

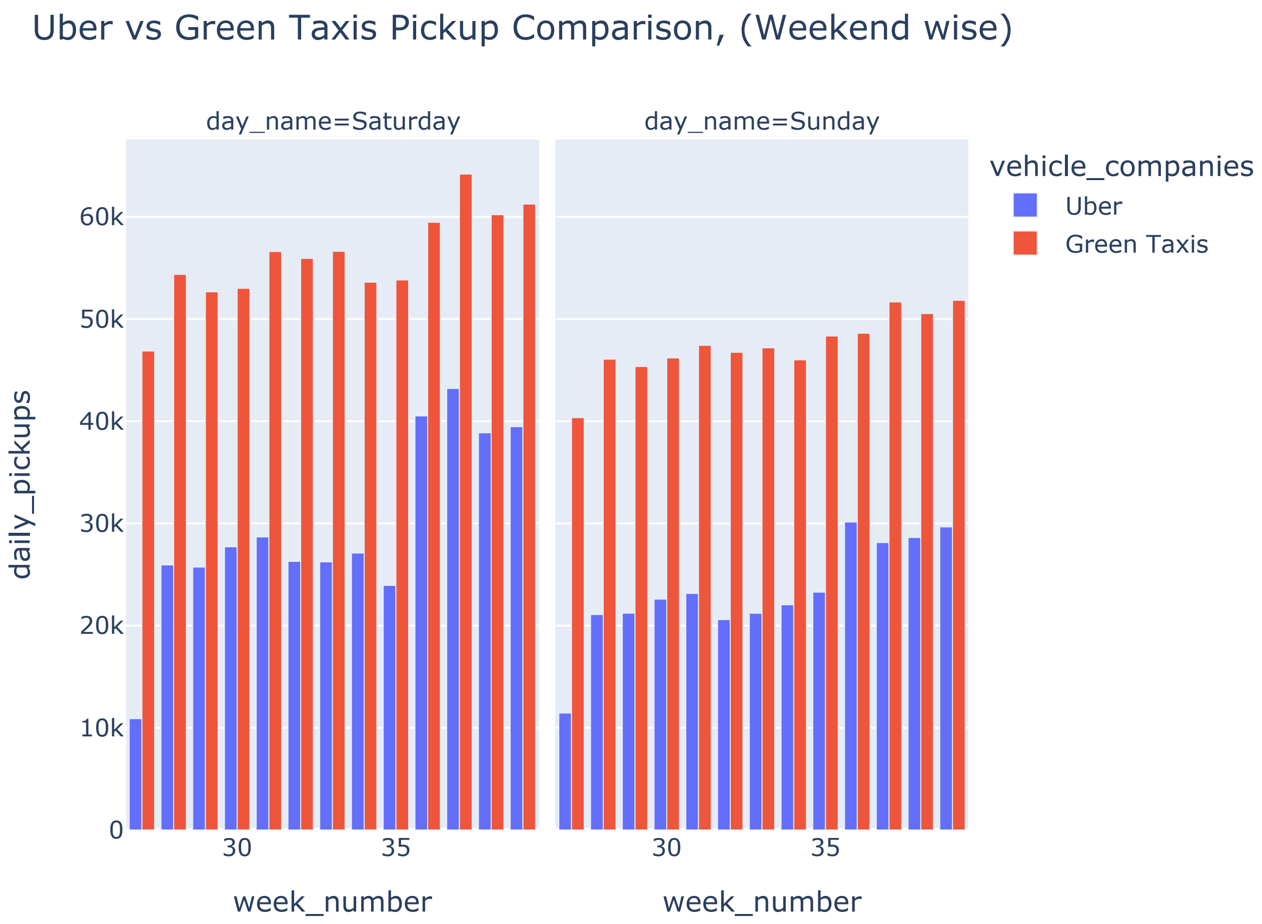
```
avg_pickups = df_stack.groupby(['vehicle_companies','day_name'])['daily_pickups'].mean()
    columns={'daily_pickups' : 'avg_pickups'})
fig = px.bar(avg_pickups, x="day_name", y="avg_pickups",
             color="vehicle_companies",
             category_orders={"day_name": ["Monday", "Tuesday","Wednesday",
                                           "Thursday", "Friday", "Saturday", "Sunday"]},
             title="Avg. Daywise vehicle Pickup ",
             barmode="group")
fig.show()
```



1. It is interesting to see that the avg. Pickup counts consistently increase for Uber and Green taxis.
2. The avg. pickup number differs moderately on the weekdays but drops considerably for Uber over the weekend.

In [23]:

```
## Uber vs Green Taxis pickup trend
uber_greentaxis = df_stack[(df_stack.vehicle_companies.isin(['Green Taxis','Uber'])) & (
fig = px.bar(uber_greentaxis, x="week_number", y="daily_pickups", color="vehicle_compani
             barmode="group",
             category_orders={"day_name": ["Saturday","Sunday"]},
             facet_col ="day_name",
            title="Uber vs Green Taxis Pickup Comparison, (Weekend wise)")
fig.show()
```

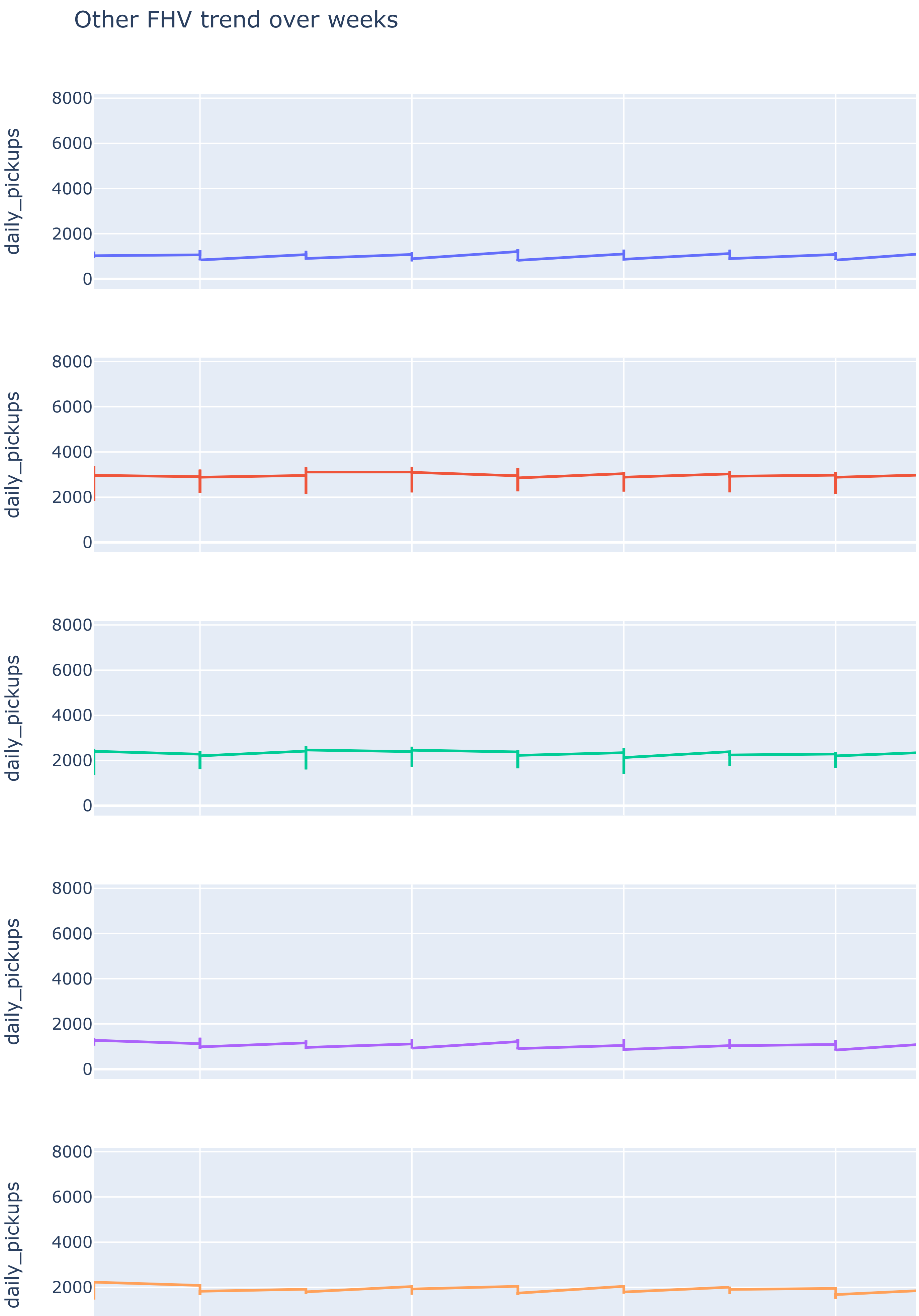


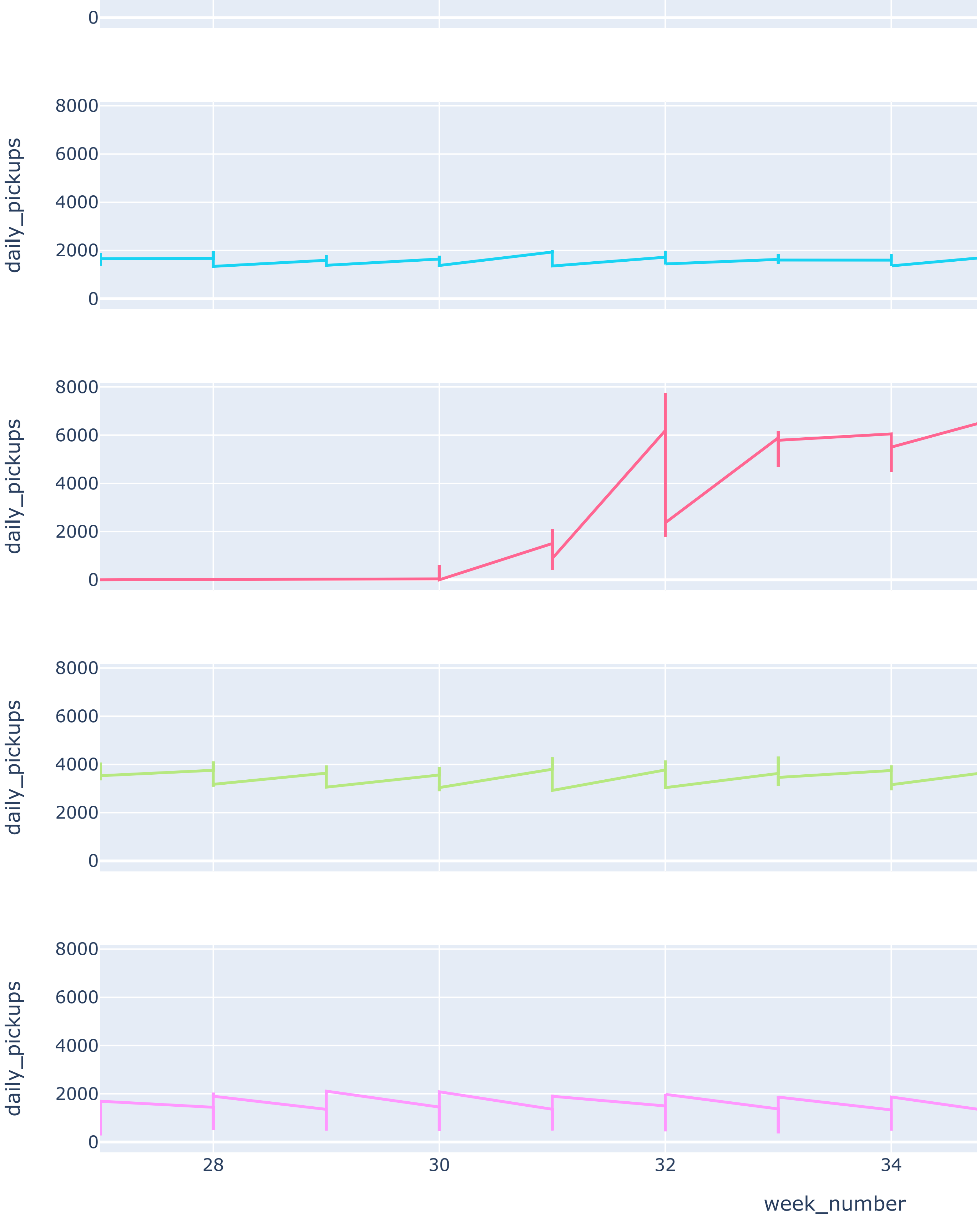
1. Saturday pickup numbers are higher than Sunday and have hit the 60K mark in the 37th week.
2. Sunday pickup numbers are consistently lower for Uber compared to Green taxis.

Part 2 : Analysing other FHV

In [24]:

```
other_fhv = df_stack[~df_stack.vehicle_companies.isin(['Green Taxis','Uber'])]
other_fhv_grouper = other_fhv.groupby(['vehicle_companies','week_number','day_name'])['d
fig = px.line(other_fhv_grouper, x="week_number", y="daily_pickups", color="vehicle_comp
             category_orders={"day_name": ["Monday","Tuesday","Wednesday","Thursday","Fr
             facet_row="vehicle_companies",
            title="Other FHV trend over weeks",
             width= 1300, height=1900)
fig.show()
```





1. Out of all 9 Other FHVs (excluding Uber and Green/Yellow taxis, Lyft has the highest pickup stats.
2. Most other vehicles have had a stable pickup trend over the last three months, but there is a slight weekly trend for Prestige and Skyling.
3. Camel and Dial 7 have fever fluctuations, suggesting that these vehicles' market share is stable.

Overall Analysis

1. Based on the current data pickup trend is shifting towards Uber. I recommend including the pricing data to understand how fares impact the pickup demand.
2. Yellow taxis hold most of the pickup market share as they are permitted to pick up from anywhere in the city. Comparing this data with other cities would be interesting to understand if this trend is reflected in other cities too. As a business strategy, we can also have a long-term plan to get these permissions for the FHV segment.
3. Saturdays have overall higher pickup numbers over Sunday, but Uber pickups drop significantly to Green taxis from Saturday to Sunday. Comparing this data with other cities would be interesting to understand if this trend is reflected in other cities too.
4. In the other FHV segment (excluding Uber and Yellow/Green taxis), Lyft data fluctuate even with a reasonable growth rate. This suggests that it is growing its market share as a new entrant. I recommend we start **collecting customer reviews** to understand what is working or not.
5. The 3-month snapshot gives a small glimpse of the pickup data for NYC. Including more data from different cities, **including the pricing data, weather data, driver attributes and geodata** , can help us understand the impact of these variables on the pickup rates.