Appendix B

Jupyter Notebook part 2 (the page number is only for this Notebook)

[1]:
```
import pandas as pd
import numpy as np
from numpy import log, sqrt
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

[2]:
```
df = pd.read_csv('data/yellow_tripdata_2016-01.csv',error_bad_lines=False)
df
```

[2]:
```
          VendorID tpep_pickup_datetime tpep_dropoff_datetime  \
0                2  2016-01-01 00:00:00   2016-01-01 00:00:00
1                2  2016-01-01 00:00:00   2016-01-01 00:00:00
2                2  2016-01-01 00:00:00   2016-01-01 00:00:00
3                2  2016-01-01 00:00:00   2016-01-01 00:00:00
4                2  2016-01-01 00:00:00   2016-01-01 00:00:00
…              …                    …                    …
10906853         2  2016-01-31 23:30:32   2016-01-31 23:38:18
10906854         1  2016-01-05 00:15:55   2016-01-05 00:16:06
10906855         1  2016-01-05 06:12:46   2016-03-19 20:45:50
10906856         1  2016-01-05 06:21:44   2016-03-28 12:54:26
10906857         1  2016-01-05 06:15:21   2016-01-05 06:15:36

          passenger_count  trip_distance  pickup_longitude  pickup_latitude  \
0                       2           1.10       -73.990372        40.734695
1                       5           4.90       -73.980782        40.729912
2                       1          10.54       -73.984550        40.679565
3                       1           4.75       -73.993469        40.718990
4                       3           1.76       -73.960625        40.781330
…                     …             …                  …                …
10906853                1           2.20       -74.003578        40.751011
10906854                1           0.00       -73.945488        40.751530
10906855                3           1.40       -73.994240        40.766586
10906856                1           2.10       -73.948067        40.776531
10906857                3           0.00       -73.960938        40.758595

          RatecodeID store_and_fwd_flag  dropoff_longitude  dropoff_latitude  \
```

```
0                  1                  N         -73.981842         40.732407
1                  1                  N         -73.944473         40.716679
2                  1                  N         -73.950272         40.788925
3                  1                  N         -73.962242         40.657333
4                  1                  N         -73.977264         40.758514
…                …                  …                  …                …
10906853           1                  N         -73.982651         40.767509
10906854           1                  N         -73.945457         40.751530
10906855           1                  N         -73.984428         40.753922
10906856           1                  N         -73.978188         40.777435
10906857           2                  N         -73.961006         40.758583

          payment_type  fare_amount  extra  mta_tax  tip_amount  tolls_amount  \
0                    2          7.5    0.5      0.5        0.00          0.00
1                    1         18.0    0.5      0.5        0.00          0.00
2                    1         33.0    0.5      0.5        0.00          0.00
3                    2         16.5    0.0      0.5        0.00          0.00
4                    2          8.0    0.0      0.5        0.00          0.00
…                  …           …    …        …          …             …
10906853             2          8.5    0.5      0.5        0.00          0.00
10906854             2          2.5    0.5      0.5        0.00          0.00
10906855             2          7.5    0.5      0.5        0.00          0.00
10906856             1         11.5    0.0      0.5        2.45          0.00
10906857             2         52.0    0.0      0.5        0.00          5.54

          improvement_surcharge  total_amount
0                           0.3          8.80
1                           0.3         19.30
2                           0.3         34.30
3                           0.3         17.30
4                           0.3          8.80
…                         …            …
10906853                    0.3          9.80
10906854                    0.3          3.80
10906855                    0.3          8.80
10906856                    0.3         14.75
10906857                    0.3         58.34

[10906858 rows x 19 columns]
```

[3]:
```
df.columns
```

[3]:
```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'pickup_longitude',
       'pickup_latitude', 'RatecodeID', 'store_and_fwd_flag',
       'dropoff_longitude', 'dropoff_latitude', 'payment_type', 'fare_amount',
       'extra', 'mta_tax', 'tip_amount', 'tolls_amount',
```

```
       'improvement_surcharge', 'total_amount'],
      dtype='object')
```

[ ]:

[4]:
```
df.dropna(inplace=True)
df['tpep_pickup_datetime'] = pd.to_datetime(df['tpep_pickup_datetime'],
                                format='%Y/%m/%d %H:%M', errors='coerce')
df['tpep_dropoff_datetime'] = pd.to_datetime(df['tpep_dropoff_datetime'],
                                format='%Y/%m/%d %H:%M', errors='coerce')

df['duration'] = (df['tpep_dropoff_datetime'] -
                  df['tpep_pickup_datetime']).dt.seconds.astype(int) /60
```

[5]:
```
df['start_hour'] =  df['tpep_pickup_datetime'].dt.hour
df['start_date'] = df['tpep_pickup_datetime'].dt.strftime('%Y-%m-%d')
```

[ ]:

[ ]:

[6]:
```
weather = pd.read_csv('data/weather_preprocessed.csv', index_col= 'time')
weather
```

[6]:

time	0	1	2	3	4	5	6	7
2016-01-01	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-02	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-03	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-04	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-05	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-06	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-07	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-08	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-09	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-10	remain	preci	remain	remain	preci	preci	preci	preci
2016-01-11	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-12	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-13	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-14	remain	remain	remain	remain	remain	remain	preci	remain
2016-01-15	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-16	remain	remain	preci	preci	preci	preci	remain	preci
2016-01-17	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-18	remain	remain	remain	remain	remain	remain	remain	preci
2016-01-19	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-20	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-21	remain	remain	remain	remain	remain	remain	remain	remain

2016-01-22	preci	remain	remain	remain	remain	remain	remain	remain
2016-01-23	preci	preci	preci	preci	preci	preci	preci	preci
2016-01-24	remain	preci	preci	remain	remain	remain	remain	remain
2016-01-25	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-26	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-27	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-28	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-29	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-30	remain	remain	remain	remain	remain	remain	remain	remain
2016-01-31	remain	remain	remain	remain	remain	remain	remain	remain

time	8	9	…	14	15	16	17	18	\
2016-01-01	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-02	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-03	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-04	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-05	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-06	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-07	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-08	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-09	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-10	preci	preci	…	remain	remain	preci	remain	remain	
2016-01-11	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-12	remain	remain	…	remain	remain	remain	preci	remain	
2016-01-13	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-14	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-15	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-16	preci	remain	…	remain	remain	remain	remain	remain	
2016-01-17	remain	remain	…	remain	preci	preci	preci	preci	
2016-01-18	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-19	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-20	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-21	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-22	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-23	preci	preci	…	preci	preci	preci	preci	preci	
2016-01-24	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-25	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-26	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-27	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-28	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-29	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-30	remain	remain	…	remain	remain	remain	remain	remain	
2016-01-31	remain	remain	…	remain	remain	remain	remain	remain	

time	19	20	21	22	23

```
2016-01-01  remain  remain  remain  remain  remain
2016-01-02  remain  remain  remain  remain  remain
2016-01-03  remain  remain  remain  remain  remain
2016-01-04  remain  remain  remain  remain  remain
2016-01-05  remain  remain  remain  remain  remain
2016-01-06  remain  remain  remain  remain  remain
2016-01-07  remain  remain  remain  remain  remain
2016-01-08  remain  remain  remain  remain  remain
2016-01-09  remain  remain   preci  remain  remain
2016-01-10  remain  remain  remain  remain  remain
2016-01-11  remain  remain  remain  remain  remain
2016-01-12  remain  remain  remain  remain  remain
2016-01-13  remain  remain  remain  remain  remain
2016-01-14  remain  remain  remain  remain  remain
2016-01-15  remain  remain  remain  remain  remain
2016-01-16  remain  remain  remain  remain  remain
2016-01-17   preci   preci  remain  remain  remain
2016-01-18  remain  remain  remain  remain  remain
2016-01-19  remain  remain  remain  remain  remain
2016-01-20  remain  remain  remain  remain  remain
2016-01-21  remain  remain  remain  remain  remain
2016-01-22  remain  remain  remain  remain   preci
2016-01-23   preci   preci   preci   preci   preci
2016-01-24  remain  remain  remain  remain  remain
2016-01-25  remain  remain  remain  remain  remain
2016-01-26  remain  remain  remain  remain  remain
2016-01-27  remain  remain  remain  remain  remain
2016-01-28  remain  remain  remain  remain  remain
2016-01-29  remain  remain  remain  remain  remain
2016-01-30  remain  remain  remain  remain  remain
2016-01-31  remain  remain  remain  remain  remain

[31 rows x 24 columns]
```

[ ]:

[7]:
```
def fill_weather(weather, date, time):
    return weather.loc[date][time]


df['weather'] = df[['start_date','start_hour']].apply(
    lambda x: fill_weather(weather,x.iloc[0],x.iloc[1]), axis=1)
```

[8]:
```
df.to_feather('data/yellow_tripdata_01_weather.feather')
```

[9]:
```
df
```

[9]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	\
0	2	2016-01-01 00:00:00	2016-01-01 00:00:00	
1	2	2016-01-01 00:00:00	2016-01-01 00:00:00	
2	2	2016-01-01 00:00:00	2016-01-01 00:00:00	
3	2	2016-01-01 00:00:00	2016-01-01 00:00:00	
4	2	2016-01-01 00:00:00	2016-01-01 00:00:00	
…	…	…	…	
10906853	2	2016-01-31 23:30:32	2016-01-31 23:38:18	
10906854	1	2016-01-05 00:15:55	2016-01-05 00:16:06	
10906855	1	2016-01-05 06:12:46	2016-03-19 20:45:50	
10906856	1	2016-01-05 06:21:44	2016-03-28 12:54:26	
10906857	1	2016-01-05 06:15:21	2016-01-05 06:15:36	

	passenger_count	trip_distance	pickup_longitude	pickup_latitude	\
0	2	1.10	-73.990372	40.734695	
1	5	4.90	-73.980782	40.729912	
2	1	10.54	-73.984550	40.679565	
3	1	4.75	-73.993469	40.718990	
4	3	1.76	-73.960625	40.781330	
…	…	…	…	…	
10906853	1	2.20	-74.003578	40.751011	
10906854	1	0.00	-73.945488	40.751530	
10906855	3	1.40	-73.994240	40.766586	
10906856	1	2.10	-73.948067	40.776531	
10906857	3	0.00	-73.960938	40.758595	

	RatecodeID	store_and_fwd_flag	dropoff_longitude	…	extra	\
0	1	N	-73.981842	…	0.5	
1	1	N	-73.944473	…	0.5	
2	1	N	-73.950272	…	0.5	
3	1	N	-73.962242	…	0.0	
4	1	N	-73.977264	…	0.0	
…	…	…	…	…	…	
10906853	1	N	-73.982651	…	0.5	
10906854	1	N	-73.945457	…	0.5	
10906855	1	N	-73.984428	…	0.5	
10906856	1	N	-73.978188	…	0.0	
10906857	2	N	-73.961006	…	0.0	

	mta_tax	tip_amount	tolls_amount	improvement_surcharge	\
0	0.5	0.00	0.00	0.3	
1	0.5	0.00	0.00	0.3	
2	0.5	0.00	0.00	0.3	
3	0.5	0.00	0.00	0.3	
4	0.5	0.00	0.00	0.3	
…	…	…	…	…	
10906853	0.5	0.00	0.00	0.3	

```
10906854      0.5        0.00          0.00                    0.3
10906855      0.5        0.00          0.00                    0.3
10906856      0.5        2.45          0.00                    0.3
10906857      0.5        0.00          5.54                    0.3

          total_amount    duration  start_hour  start_date  weather
0                 8.80    0.000000           0  2016-01-01   remain
1                19.30    0.000000           0  2016-01-01   remain
2                34.30    0.000000           0  2016-01-01   remain
3                17.30    0.000000           0  2016-01-01   remain
4                 8.80    0.000000           0  2016-01-01   remain
…                 …         …          …          …      …
10906853          9.80    7.766667          23  2016-01-31   remain
10906854          3.80    0.183333           0  2016-01-05   remain
10906855          8.80  873.066667           6  2016-01-05   remain
10906856         14.75  392.700000           6  2016-01-05   remain
10906857         58.34    0.250000           6  2016-01-05   remain

[10906858 rows x 23 columns]
```