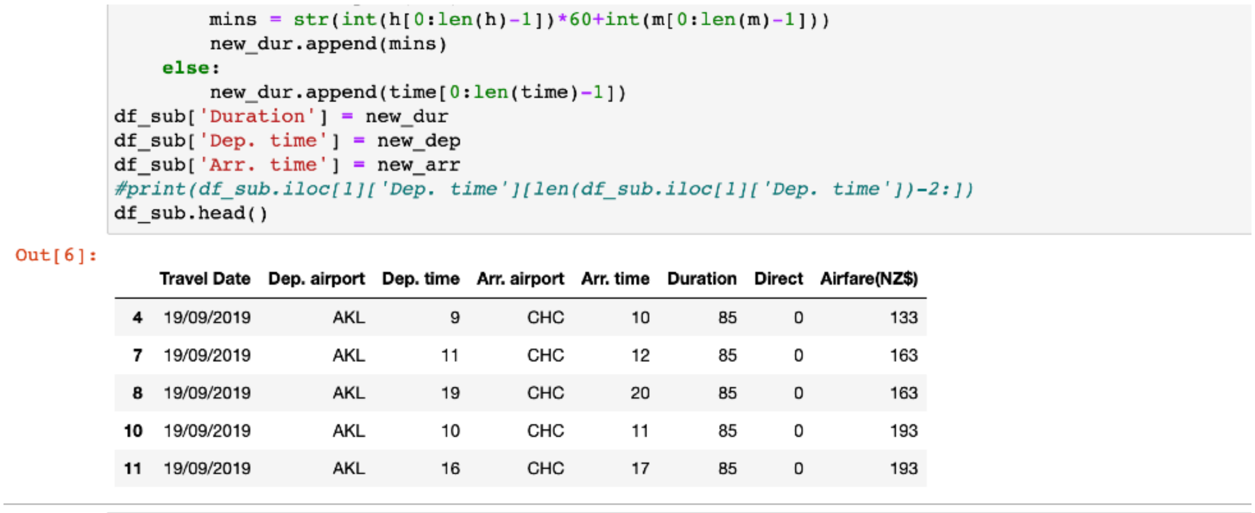
CSE400 Assignment 2 report

Huahao Shang
hushang@syr.edu
SUID: hushang

1.
(a) 162833 before the filter and 157431 rows after filter

(b) Depart Airport and Arrive Airport have NaN values, 157407 rows left after drop those.

(c)
Depart Airports: [‘AKL' 'CHC' 'WLG' ‘ZQN'] 4 types
Arrival Airports: [‘CHC' 'DUD' 'NPE' 'NPL' 'NSN' 'PMR' 'WLG' 'ZQN' ‘AKL'] 9 types
(d)

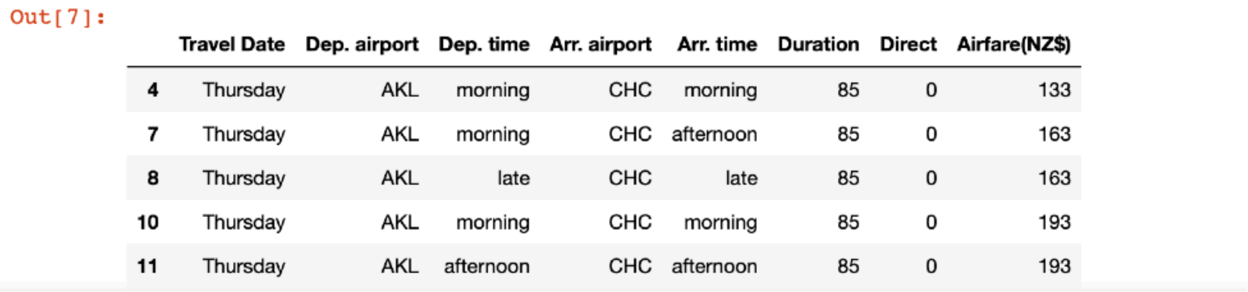


```
        mins = str(int(h[0:len(h)-1])*60+int(m[0:len(m)-1]))
        new_dur.append(mins)
    else:
        new_dur.append(time[0:len(time)-1])
df_sub['Duration'] = new_dur
df_sub['Dep. time'] = new_dep
df_sub['Arr. time'] = new_arr
#print(df_sub.iloc[1]['Dep. time'][len(df_sub.iloc[1]['Dep. time'])-2:])
df_sub.head()
```

Out[6]:

	Travel Date	Dep. airport	Dep. time	Arr. airport	Arr. time	Duration	Direct	Airfare(NZ$)
4	19/09/2019	AKL	9	CHC	10	85	0	133
7	19/09/2019	AKL	11	CHC	12	85	0	163
8	19/09/2019	AKL	19	CHC	20	85	0	163
10	19/09/2019	AKL	10	CHC	11	85	0	193
11	19/09/2019	AKL	16	CHC	17	85	0	193

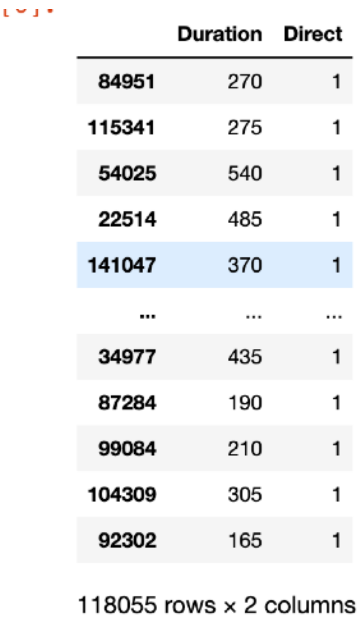
(e)



Out[7]:

	Travel Date	Dep. airport	Dep. time	Arr. airport	Arr. time	Duration	Direct	Airfare(NZ$)
4	Thursday	AKL	morning	CHC	morning	85	0	133
7	Thursday	AKL	morning	CHC	afternoon	85	0	163
8	Thursday	AKL	late	CHC	late	85	0	163
10	Thursday	AKL	morning	CHC	morning	85	0	193
11	Thursday	AKL	afternoon	CHC	afternoon	85	0	193

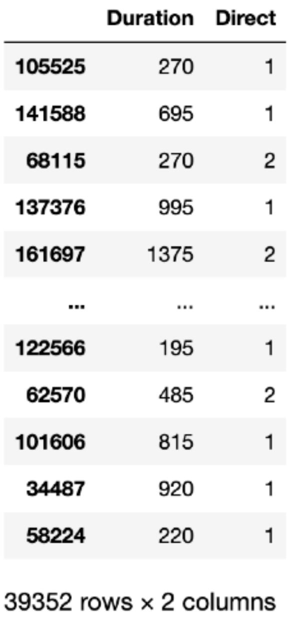
2.
(a)
TrainX



	Duration	Direct
84951	270	1
115341	275	1
54025	540	1
22514	485	1
141047	370	1
...	...	...
34977	435	1
87284	190	1
99084	210	1
104309	305	1
92302	165	1

118055 rows × 2 columns

Test X



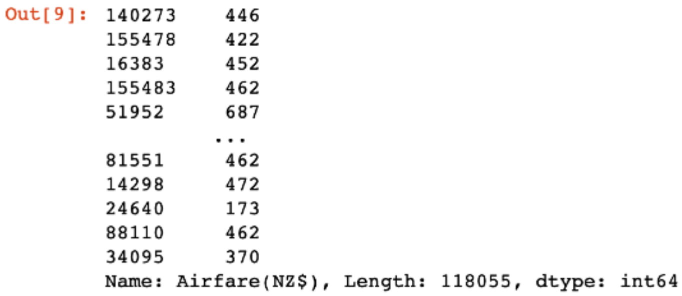
	Duration	Direct
105525	270	1
141588	695	1
68115	270	2
137376	995	1
161697	1375	2
...	...	...
122566	195	1
62570	485	2
101606	815	1
34487	920	1
58224	220	1

39352 rows × 2 columns

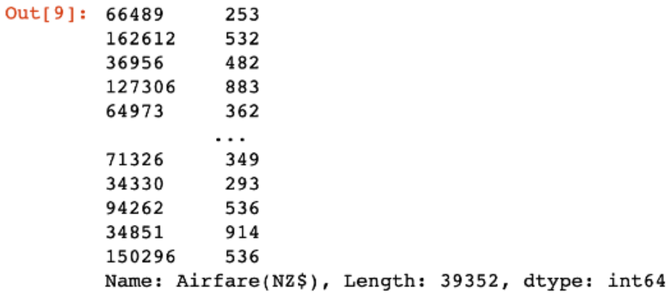
Train: 118055 rows.
First index: 84951

Test: 39352 rows.
Frist index: 105525

TrainY and TestY have the same rows and index of TrainX and TestX



```
Out[9]: 140273    446
        155478    422
        16383     452
        155483    462
        51952     687
                 ...
        81551     462
        14298     472
        24640     173
        88110     462
        34095     370
        Name: Airfare(NZ$), Length: 118055, dtype: int64
```

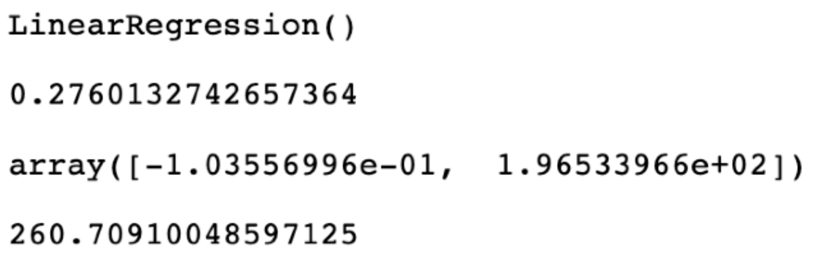


```
Out[9]: 66489     253
        162612    532
        36956     482
        127306    883
        64973     362
                 ...
        71326     349
        34330     293
        94262     536
        34851     914
        150296    536
        Name: Airfare(NZ$), Length: 39352, dtype: int64
```

(b)

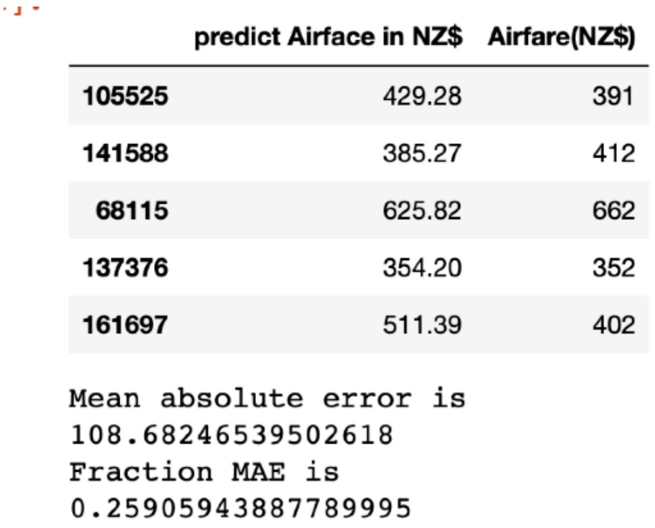
R^2: 0.276   Beta 1: -0.1035 Beta 2: 196.53 Beta 0: 260.70

Beta 1 shows Duration and Airfare have a little negative relatively, might because shorter flight tends to be expensive.
Beta 2 shows Direct has strong positive relativity with Airfare, might because Direct only contain 1, 2, 3, or 0, but Airfare are large.



```
LinearRegression()

0.2760132742657364

array([-1.03556996e-01,  1.96533966e+02])

260.70910048597125
```

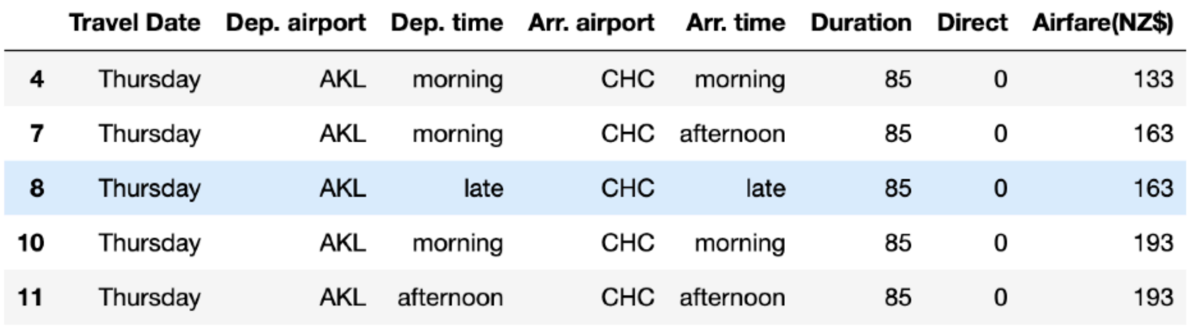
(c) MAE: 108.68 MAE to the average of AirFare: 0.259
    The prediction value and the real value are pretty good fit. In the below chart, 3 of 6 predictions are almost the same, and 5 of 6 different value less than 100.



	predict Airface in NZ$	Airfare(NZ$)
105525	429.28	391
141588	385.27	412
68115	625.82	662
137376	354.20	352
161697	511.39	402

```
Mean absolute error is
108.68246539502618
Fraction MAE is
0.25905943887789995
```

3.
(a)



	Travel Date	Dep. airport	Dep. time	Arr. airport	Arr. time	Duration	Direct	Airfare(NZ$)
4	Thursday	AKL	morning	CHC	morning	85	0	133
7	Thursday	AKL	morning	CHC	afternoon	85	0	163
8	Thursday	AKL	late	CHC	late	85	0	163
10	Thursday	AKL	morning	CHC	morning	85	0	193
11	Thursday	AKL	afternoon	CHC	afternoon	85	0	193

(b)

25 Columns

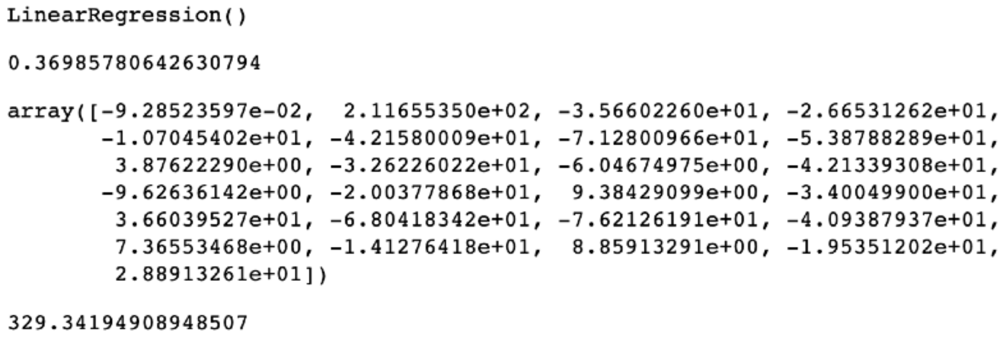
Durations and Direct are non-categorial column 2, week days(Mon, Tue….)7, Airport Names 13 and (early, late, morning, afternoon) 4 are all categorial values and placed into binary representation.

In total 26 columns, but mine data set is missing Friday, I don't know why, only display 25 columns

(c)

Train rows: 118055.   Test rows: 39352
First row index: 82184.   First row index: 102055

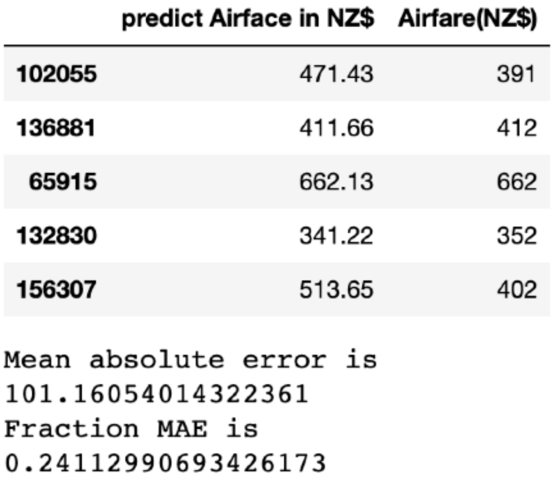
(d)



```
LinearRegression()

0.36985780642630794

array([-9.28523597e-02,  2.11655350e+02, -3.56602260e+01, -2.66531262e+01,
       -1.07045402e+01, -4.21580009e+01, -7.12800966e+01, -5.38788289e+01,
        3.87622290e+00, -3.26226022e+01, -6.04674975e+00, -4.21339308e+01,
       -9.62636142e+00, -2.00377868e+01,  9.38429099e+00, -3.40049900e+01,
        3.66039527e+01, -6.80418342e+01, -7.62126191e+01, -4.09387937e+01,
        7.36553468e+00, -1.41276418e+01,  8.85913291e+00, -1.95351202e+01,
        2.88913261e+01])

329.34194908948507
```

R^2: 0.36   Beta values are in the array follow from beta 1 to the last beta 25 (should have beta 26, but missing Friday)
Beta 3 to Beta 6 are coefficients for day of the week, they are pretty much the same relativity, all negative but various a bit in value, to the Airfare.

(e)



	predict Airface in NZ$	Airfare(NZ$)
102055	471.43	391
136881	411.66	412
65915	662.13	662
132830	341.22	352
156307	513.65	402

```
Mean absolute error is
101.16054014322361
Fraction MAE is
0.24112990693426173
```

MAE: 101.16
MAE to average Airfare: 0.241.
The prediction is more accurate than the previous one. The MAE value is decreased, means the prediction values are more close to the real value.
But the improve is not that significant in my opinion, the MAE still higher than 100. I think the added variables are not that close related to the Airfare. Flight duration and stops may affect the price more than weekdays, depart, arrive times and locations.