ST2195 - Programming for Data Science

(Coursework Report – Python and R)

UOL Student ID Number - 200643838

Page count – 10 (excluding cover page and table of contents)

Contents

Introduction

This report is based on planes that have been flying over different locations overtime and the delays that it has faced when departing from the Origin and arriving at the Destination airports within the United States of America(USA).

The “2006” and “2007” year datasets obtained from the Harvard Website were seen to be the cleanest datasets and also included the highest amount of observations out of all datasets provided since a lot of data is needed especially when creating machine learning models, hence those were chosen to be investigated in both R and Python programming languages. In addition, other .csv files on Airports and plane data were also used when answering certain questions.

This report is split into 6 parts:

1. Data Cleaning process that was done to make data more meaningful and easier to investigate.
2. Finding the best time of Day , Day of week and time of year to fly to reduce delays.
3. An investigation on whether older planes suffer more delays.
4. Checking how the number of people flying between different locations change over time.
5. Checking whether cascading delays in one airport create delays in other airports.
6. Building a model that can predict delays.

The graphics and tables made in R and Python that were helpful in deducing the final conclusion are given under each part.

Data Cleaning Process

The datasets have to be cleaned to make sure it is meaningful before we go ahead with wrangling and visualizing the data provided to us. Since the columns of both the 2006 and 2007 year datasets provided to us were the same, they were combined to form a single dataset. The duplicated rows from all the datasets (combined dataset, airports and plane data) were removed. Only missing values from the datasets other than the combined dataset were removed in the data cleaning process but the missing values in combined dataset will be removed where necessary when doing the questions. The datatypes of the column variables were checked and were seen to be in the appropriate datatypes hence no changes were made. The Scheduled departure times(CRSDepTime) of each flight has been used for questions where time of the flight is needed. The delays were taken as it is when answering the questions (delays less than 0 means that the plane has departed early, delays equal to 0 means that the plane has departed on time and delays greater than zero means there has been a delay).

Best Time of day, day of week and time of year to minimize delays?

1)**Best Day Of week**

Missing values in the arrival and departure delays from the combined dataset were removed before answering this question because missing values might mislead to an invalid final conclusion. Weekdays were grouped by the means of Arrival delay and Departure delays of the planes to show the average arrival and departure delays per weekday since averages tend to give you a general idea about the data. This is shown in the tables given below,

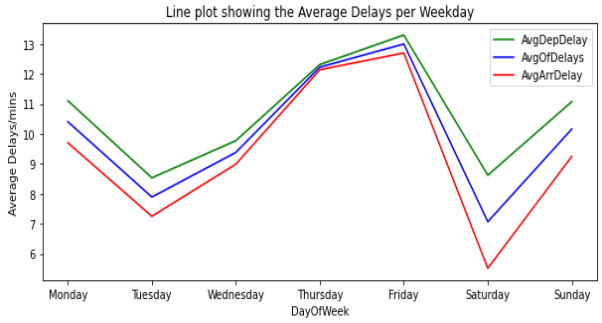
DayOfWeek	AvgArrDelay	AvgDepDelay
1	9.701697	11.106484
2	7.248419	8.532392
3	8.987949	9.771903
4	12.141546	12.314958
5	12.701592	13.298945
6	5.513338	8.620799
7	9.247210	11.074628

- Dayofweek groupby Table in Python-

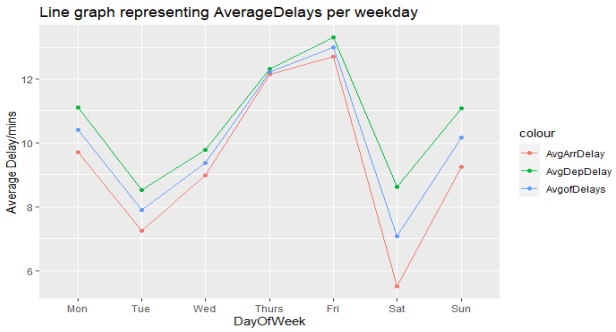
DayOfWeek <int>	AvgArrDelay <dbl>	AvgDepDelay <dbl>	AvgofDelays <dbl>
1	9.701697	11.106484	10.404091
2	7.248419	8.532392	7.890405
3	8.987949	9.771903	9.379926
4	12.141546	12.314958	12.228252
5	12.701592	13.298945	13.000268
6	5.513338	8.620799	7.067069
7	9.247210	11.074628	10.160919

-Dayofweek groupby Table in R-

A line plot was then plotted to find the best weekday to minimize delays.



-Dayofweek line plot in Python-



-Dayofweek line plot in R-

Average of the average arrival and average departure delays was found and plotted in the line graph to 8show a compromise between arrival and departure delays(blue line). From the above line plot, it is clearly visible that Saturday has the least average delays out of all weekdays hence “Saturday” is the best day of week to minimize delays. All the 3 lines show similar trend but it seems like average departure delays are larger than the average arrival delays in all days of week.

2)**Best Time of year**

The best time of year was taken to be the best month of the year when answering this question. The months were grouped by the average arrival and departure delays and made into a table , this is shown below.

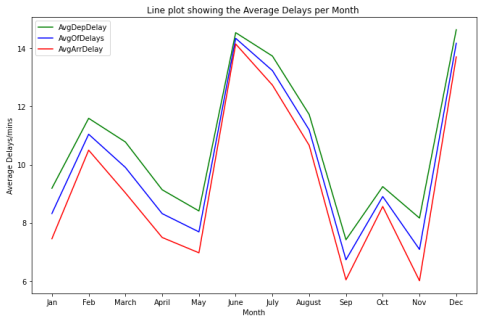
Month	AvgArrDelay	AvgDepDelay
1	7.447214	9.186764
2	10.498488	11.595069
3	9.027273	10.779701
4	7.491780	9.134426
5	6.964127	8.404951
6	14.152992	14.544010
7	12.737077	13.737311
8	10.662705	11.724292
9	6.034477	7.413950
10	8.559238	9.242656
11	6.002441	8.160109
12	13.709195	14.645250

-Month groupby Table in Python-

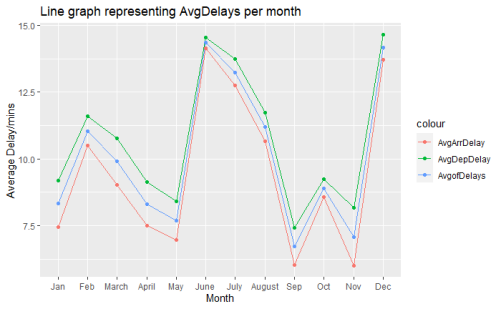
Month	AvgArrDelay	AvgDepDelay	AvgofDelays
1	7.447214	9.186764	8.316989
2	10.498488	11.595069	11.046778
3	9.027273	10.779701	9.903487
4	7.491780	9.134426	8.313103
5	6.964127	8.404951	7.684539
6	14.152992	14.544010	14.348501
7	12.737077	13.737311	13.237194
8	10.662705	11.724292	11.193498
9	6.034477	7.413950	6.724214
10	8.559238	9.242656	8.900947
11	6.002441	8.160109	7.081275
12	13.709195	14.645250	14.177222

-Month groupby Table in R-

A line plot was visualized to find the trend in average delays over months. Average of average arrival and average departure delays was found and plotted in the line graph. That average was used to find the best month to reduce delays.



-Month line plot in python-



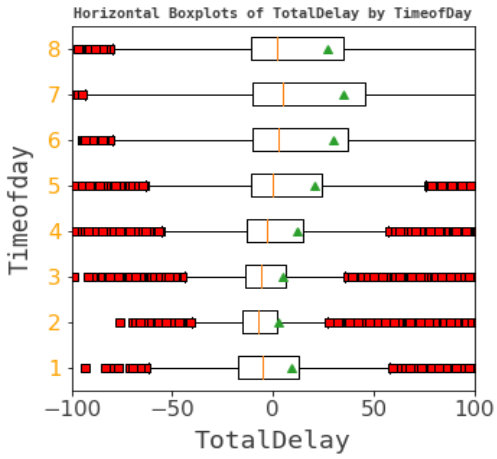
-Month line plot in R-

From the above line plot, it is visible that the blue line which represents the average of both the delays has a minimum in September hence we can conclude that “September” is the best time of the year to travel by airplanes to minimize delays. The average departure delays can again be seen to be greater than average arrival delay when grouped by months.

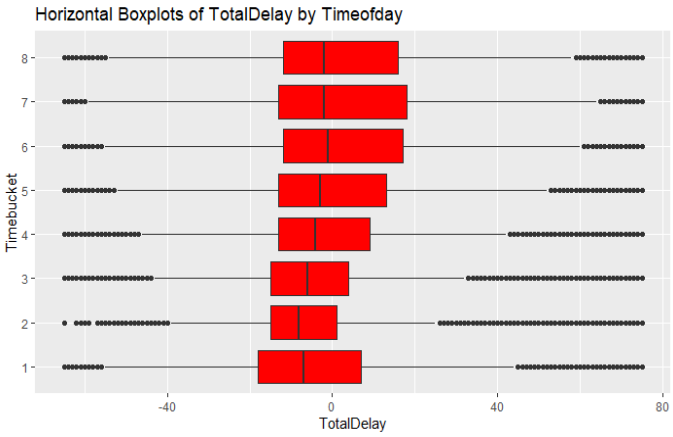
3)**Best Time of day**

In this part of the question, the delay was considered to be the sum of arrival delay and departure delays and was seen as the total delay of the planes. The Time of day was split into 8 time buckets, with each time bucket having a 3 hour time difference, starting from 00:00 to 03:00 ,03:00 to 06:00 and goes until 21:00 to 24:00.

Side by side boxplots were plotted for the total delay of planes with each boxplot in the graph below numbered from 1 to 8 representing the boxplot of total delay of each time bucket.



- Side by side boxplots in Python-



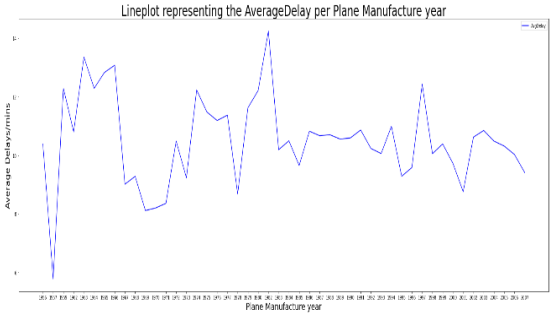
-Side by side boxplots in R-

It is visible from the boxplot diagram above that there a lot of outliers in the dataset hence median would be preferred more than the mean when comparing since median isn't affected by outliers but mean is affected. The median is shown by the middle line of the boxplots (gold line in boxplot from Python and black line in boxplot from R) and that is used for the comparison. The median of total delay of planes in the time bucket "2" and "3" seem to be really close to each other but the median of total delay in time bucket "2" appears to be slightly lower. Hence the time bucket "2" which represents the time interval [03:00 , 06:00) is the best time of day to minimize the delays of planes.

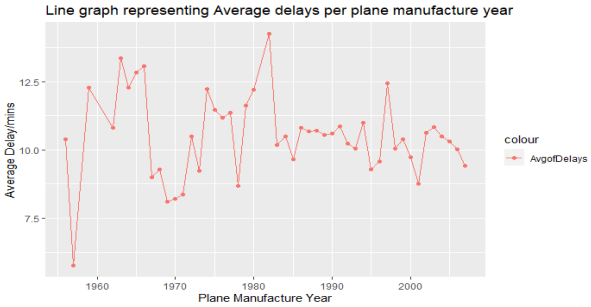
Do Older Planes suffer more delays?

Age of the plane is taken to be associated with the manufacture year of the planes. Since the manufacturer year of the plane is needed to answer this question, the dataset on planes is merged with the combined dataset. Missing values from Arrival and Departure delays are removed from the merged dataset. The manufacture years of the planes are grouped by the averages of Departure and Arrival delay. The manufacture years that made no sense were removed from the grouped table. In this question, delays were taken to be the average of the average departure delay and average arrival delay.

A line plot was then constructed between average delay and the plane manufacture year to see the trend over the years.



-Line plot in Python-



-Line plot in R-

The above line plot shows us that there are fluctuations in the line as the plane manufacture year increases but it doesn't necessarily show a trend (increasing or decreasing). There is a large drop in average of delay for the manufacture year " 1957" and a large rise at "1982". There isn't necessarily much information to show that older planes do suffer more delays from the visualization.

Correlation between Avgofdelays and the plane manufacture year is found below,

```
array([[ 1.        , -0.07109128],
       [-0.07109128,  1.        ]])
```

- Correlation in Python-

```
-0.07109128
```

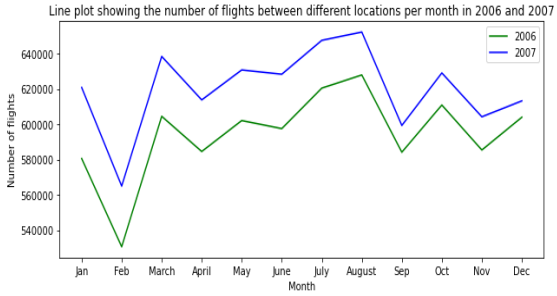
-Correlation in R-

This conclusion is also supported by the fact that the Pearson correlation between average of delays and the manufacture year is -0.07109128 which is quite close to 0 , allowing us to state that there is no supporting evidence to suggest that "Older planes suffer more delays". Since the Pearson correlation is almost 0, it indicates there is almost no linear relationship between the Plane Manufacture Year and the delay of the plane and hence the analysis shows that "Older Planes do not suffer more delays".

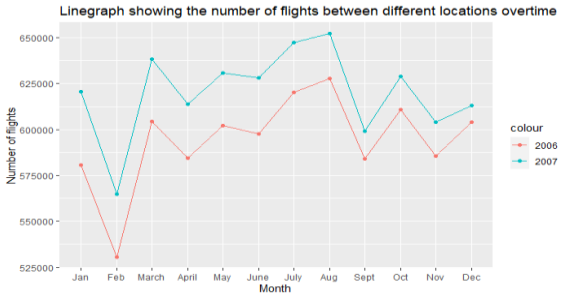
How does the number of people flying between different locations change over time?

Since we don't have any data on passengers, the number of flights is taken as a proxy (substitute) to the number of people in this question. The dataset on airports and combined dataset were merged to form a merged dataset. The merged dataset was then split into 2 datasets of which one contained the 2006 year data and the other one containing the 2007 data. Months were then grouped by the count of the flights in both the split datasets. The count of flights was used instead of the average since the investigation is regarding the number of people.

Two line plots were then plotted in the same graph to show the relationship between number of people and the Months.



- Line plot in Python-



-Line plot in R-

The line plots above indicate that as months go by, the number of people travelling between different locations rise generally. August seems to have the highest number of people travelling between different locations in both the 2006 and 2007 years, this maybe because it is a holiday period and hence more people travel by planes to visit places.

In addition, the routes taken by planes in years "2006" and "2007" which had the highest number of people travelling were too found for further analysis. This was done by transferring all necessary information into tables.

	Dest_state	Origin_state	Number of people
94	CA	CA	356304
1126	TX	TX	273782
281	HI	HI	85423
54	AZ	CA	71766
93	CA	AZ	71690
...	...	...	...
575	MI	OK	1
31	AR	AR	1
1208	VA	WV	1
680	MT	OR	1
1292	WY	WY	1

-2006 Year table in python-

Origin_state:Dest_state <fctr>	Number of people <int>
CA:CA	356304
TX:TX	273782
HI:HI	85423
CA:AZ	71766
AZ:CA	71690
NV:CA	66321
CA:NV	66225
CA:FL	64914
FL:CA	64490
TX:CA	54871

-2006 Year table in R-

	Dest_state	Origin_state	Number of people
98	CA	CA	369942
1178	TX	TX	254877
295	HI	HI	104817
97	CA	AZ	72653
59	AZ	CA	72051
...	...	...	...
481	LA	LA	1
1326	WI	MT	1
413	IN	SD	1
412	IN	SC	1
580	ME	PA	1

-2007 Year table in python-

Origin_state:Dest_state <fctr>	Number of people <int>
CA:CA	369942
TX:TX	254877
HI:HI	104817
AZ:CA	72653
CA:AZ	72051
CA:NV	69075
CA:FL	68805
FL:CA	68795
NV:CA	68626
TX:CA	57359

-2007 Year table in R-

States were chosen to show the routes(locations) in these questions. It can be seen that top 5 routes with the greatest number of people travelling in 2006 are CA:CA, TX:TX, HI:HI, CA:AZ and AZ:CA and the top 5 routes in 2007 are CA:CA, TX:TX, HI:HI ,AZ:CA and CA:AZ. The top 5 routes in both 2006 and 2007 are identical except for AZ:CA and CA:AZ switching places. This indicates that these are tourist attracting

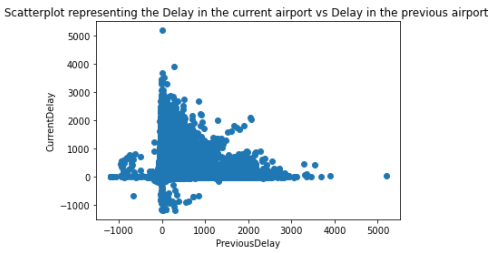
states and it is safe to say that increasing the ticket price for flights in these states will increase the profitability of the airport company. There has been a decrease in the number of people travelling in the TX:TX route. It is evident that there have been few routes where there has been only one flight.

When we take all locations as a whole, there has been an increase in the number of people travelling (or number of flights) between different locations, this is seen in the line plot.

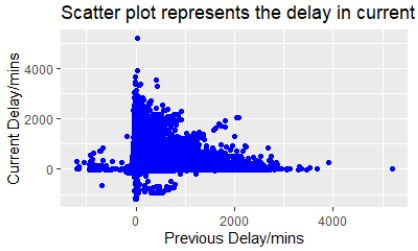
Can you detect Cascading delays in one airport create delays in others?

The delay in this question was taken to be the sum of both the arrival and departure delays of the planes. Each tail number of the flight was tracked and grouped by total delay of the plane and the Scheduled departure times of the flights were too ordered in an ascending order to make it a continuous timeline. Subsequently, the delays of the same plane in the previous airport was found. The delays in the current airport were taken to be equal to the total delay of each plane.

A scatter plot was then plotted to show the relationship between the delays in the current airport and the delays in the previous airport of the planes.



-Scatter plot in Python-



-Scatter plot in R-

It can be seen that the relationship between the delays in the current airport and the delays in the previous airport of the planes is not quite visible from the scatter plot since many points seem to be cluttered around each other.

To explore more about whether the cascading delays in one airport create delays in others, a crosstabulation is constructed to check the relationship between delay in previous airport and delay in current airport. Since crosstabs work only for categorical variables, the delay in the current airport and delays in the previous airport are changed to 1s and 0s where 1s indicate there was a delay and 0s indicate that there was no delay.

```
                Current Delay
                            0         1
Previous Delay
0               0.6859589 0.3140411
1               0.3588346 0.6411654
```

-Cross tab in R-

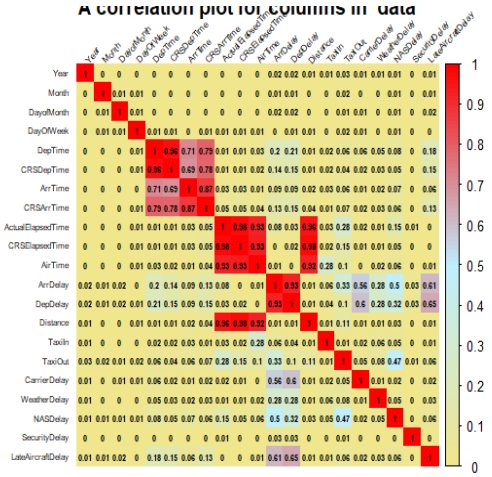
Current Delay	0	1
Previous Delay		
0	0.686008	0.313992
1	0.358896	0.641104

-Cross tab in Python-

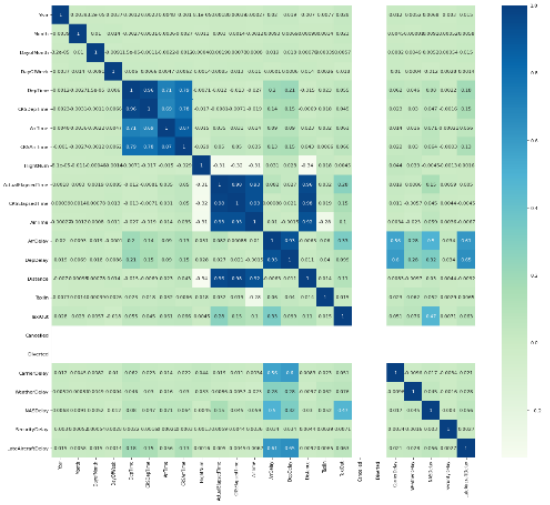
The crosstabs above indicate the probability of an event happening. It is clearly visible that the probability that there is a delay in the current airport given that there is a delay in the previous airport (0.6411)is greater than the probability that there is a delay in the current airport given that there is no delay in the previous airport(0.314). Hence, we can conclude that there is evidence from the crosstab that cascading delays in one airport create delays in other airports

Constructing a model to predict delays using the available variables.

Since the question asks us to build a model to predict delays, a multiple linear regression model is constructed for this question. The delay that is going to be predicted is the Arrival delay in the regression model. The correlation (linear relationship)between the other available variables is found using the below correlation plot.



-Correlation heatmap in R-



-Correlation heatmap in Python-

After referring the correlation heatmap, the variables that are most correlated with the Arrival delay of the plane and some variables that had some connection with the arrival delay of the planes like the day of week, month and scheduled departure time (some weekdays, some months and few time intervals had greater delays than some others as seen in the previous questions) were taken to be the explanatory variables for the model. The plane manufacture year wasn't taken as a variable since we found no evidence previously to suggest that older planes do suffer more delays.

Arrival delay is predicted assuming that the delay is predicted at the Destination airport hence the delay in departure of the plane and the late aircraft delay will also be known and will be used as explanatory variables. Weather, NAS , carrier and security delays aren't included in the model since they won't be known until the plane lands in the destination airport.

```
Call:
lm(formula = ArrDelay ~ Month + DayofMonth + DayOfWeek + CRSDepTime +
    DepDelay + LateAircraftDelay, data = scaled_train)

Residuals:
     Min       1Q   Median       3Q      Max
-151.88    -7.60    -1.61     5.28 1489.35

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.445874   0.004521 2089.38   <2e-16 ***
Month             0.111143   0.004522   24.58   <2e-16 ***
DayofMonth        0.064248   0.004522   14.21   <2e-16 ***
DayOfWeek        -0.305388   0.004522  -67.54   <2e-16 ***
CRSDepTime       -0.179946   0.004584  -39.26   <2e-16 ***
DepDelay         34.736856   0.005959 5829.58   <2e-16 ***
LateAircraftDelay  0.716055   0.005955  120.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.29 on 9995332 degrees of freedom
Multiple R-squared:  0.8583,    Adjusted R-squared:  0.8583
F-statistic: 1.009e+07 on 6 and 9995332 DF,  p-value: < 2.2e-16
```

-Model evaluation in R-

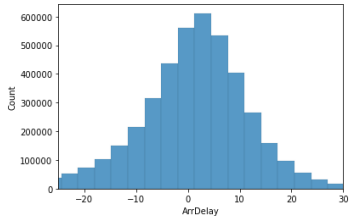
```
MSE:  205.8311065380408
MAE:  9.286270201254785
RMSE:  14.346815205405024
R squared 0.86

"The mean squared error for this model is 209.382780763958"
"The root mean squared error for this model is 14.4700649882424"
```

-Model evaluation in Python-

It is observed from the output from R, that the explanatory variables are highly significant(3 stars).

Since the adjusted R squared of this model is 0.86 which is high, we can say that this model is a good fit for predicting delays(Arrival delay of the plane) and that 86% of the variation of the target variable (Arrival delay) is explained by the variation of the explanatory variables. The Root Mean Squared error is equal to 14.35 which is moderate amount of error since the range of the Arrival delays of planes is quite high.



-Histogram of ArrDelay in Python-

-Histogram of the residuals in R-

The histogram of the target variable in Python seems to be symmetric and a bell curve (has a normal distribution) and the histogram of the residuals in R seems to be symmetric and has a normal distribution. This shows that the errors are normally distributed since the target variable being normally distributed implies that the errors are normal.

From the evaluation of the model, the multiple linear regression model seems to be a really good model for the prediction of the delays of the planes. Further improvement in the model can be done by removing certain variables to improve the R squared of the model but since the R squared for this model is quite high, this wasn't done.