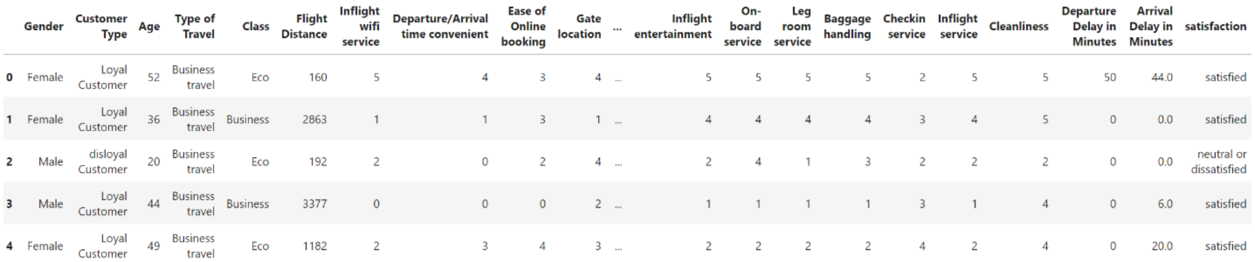
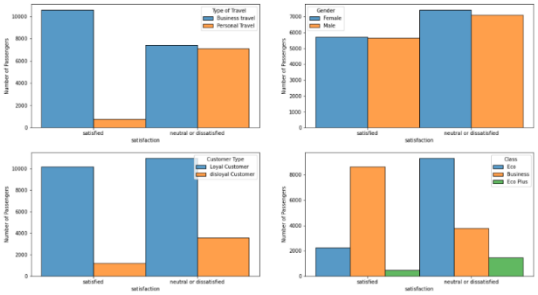
Term Project – Tianyi Liang

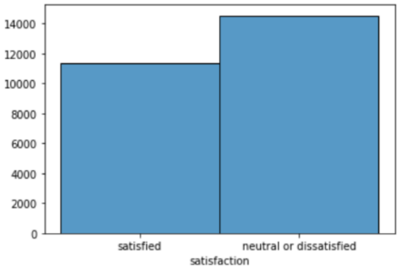
The dataset I chose records airline passenger satisfaction with 22 different attributes. The full dataset is 15.23 MB dividing to a 12.19MB train set and a 3.04 MB test set. I intend to try to correlate these 22 attributes with the satisfaction level (as a classification factor). The output column (satisfaction) is recorded with two attributes, “satisfied” or “unsatisfied”. By turning this column into a binary column with 0s and 1s, the model can use standard supervised learning classification methods.



	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	Female	Loyal Customer	52	Business travel	Eco	160	5	4	3	4	...	5	5	5	5	2	5	5	50	44.0	satisfied
1	Female	Loyal Customer	36	Business travel	Business	2863	1	1	3	1	...	4	4	4	4	3	4	5	0	0.0	satisfied
2	Male	disloyal Customer	20	Business travel	Eco	192	2	0	2	4	...	2	4	1	3	2	2	2	0	0.0	neutral or dissatisfied
3	Male	Loyal Customer	44	Business travel	Business	3377	0	0	0	2	...	1	1	1	1	3	1	4	0	6.0	satisfied
4	Female	Loyal Customer	49	Business travel	Eco	1182	2	3	4	3	...	2	2	2	2	4	2	4	0	20.0	satisfied

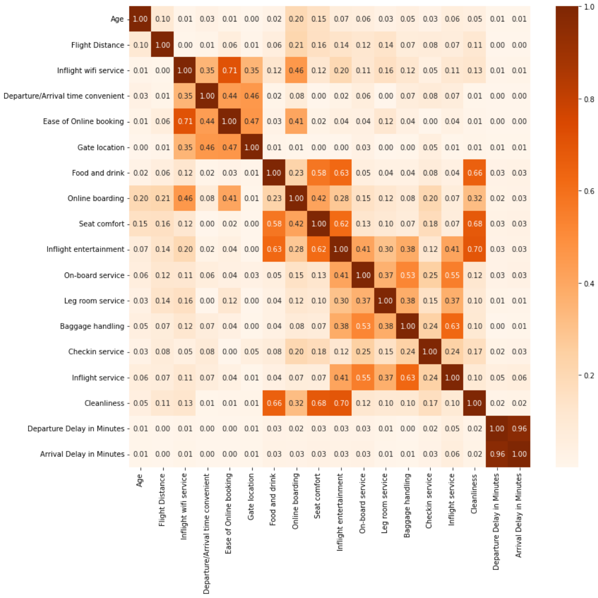
To get an overview of the dataset, I analyzed the distribution of the satisfaction levels within each attribute. The overall distribution of the two classes is balanced with approximately 20% of the difference in amount, while it is not so balanced when statistic is done inside each attribute separately. The number of people having business trip is larger than the number of people having personal trip. People having business trip tend to give more positive feedback while the most portion of the people with personal purpose give negative feedback. This attribute will be an important explanatory variable when predicting satisfaction level. The number of Female and Male is well-balanced. The ratio of positive and negative feedback differed by gender was not significant. Disloyal customers are more likely to give negative feedbacks while loyal customers give evenly distributed pos/neg feedbacks. Feedback given by people in different classes showed a completely different distribution. The sample size for Eco class and Business class is about the same, while the sample size for Eco Plus class is relatively small.

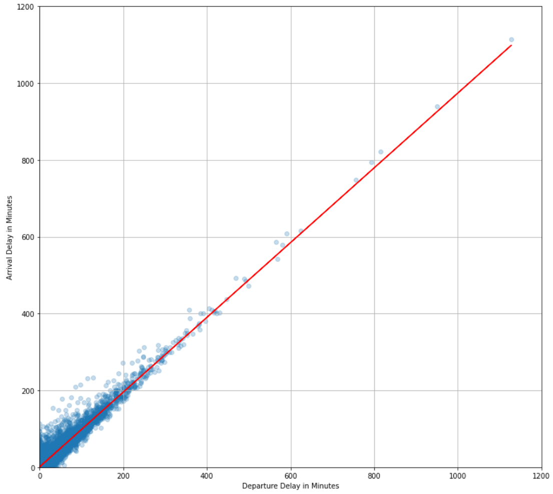


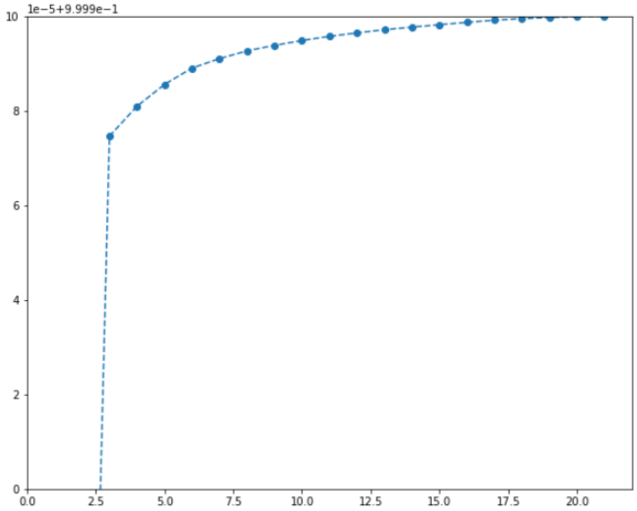


To further implement model training, one must make sure that attributes are linearly independent with each other’s. By plotting heatmap of correlation coefficient, the columns "Departure Delay in Minutes" and "Arrival Delay in Minutes" do not meet the requirement with the correlation coefficient of 0.96.

The explanatory/independent variable is apparently the "Departure Delay in Minutes" column. Therefore, the dependent variable column ("Arrival Delay in Minutes") will be dropped when training the model.



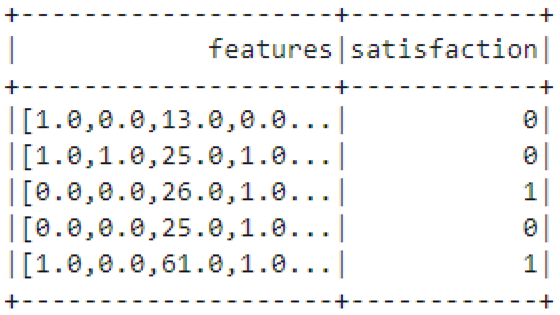




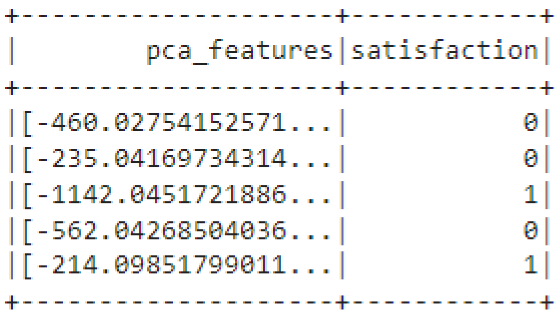
PCA Analysis

After dropping out the unnecessary columns and NAs, I performed a PCA analysis to see the relationship of feature number and variance explained.

By using Vector Assembler, all useful features shall be combined into one feature vector. With passing in the feature vector to the PCA transformation, it produces a new column recording the processed feature vector using principal component analysis.



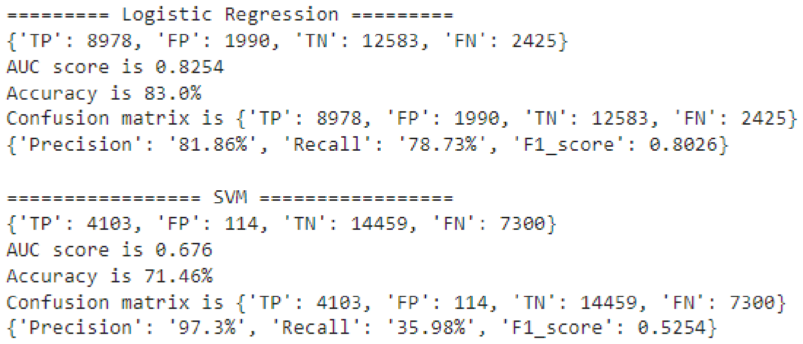
```
+--------------------+------------+
|            features|satisfaction|
+--------------------+------------+
|[1.0,0.0,13.0,0.0...|           0|
|[1.0,1.0,25.0,1.0...|           0|
|[0.0,0.0,26.0,1.0...|           1|
|[0.0,0.0,25.0,1.0...|           0|
|[1.0,0.0,61.0,1.0...|           1|
+--------------------+------------+
```



```
+--------------------+------------+
|        pca_features|satisfaction|
+--------------------+------------+
|[-460.02754152571...|           0|
|[-235.04169734314...|           0|
|[-1142.0451721886...|           1|
|[-562.04268504036...|           0|
|[-214.09851799011...|           1|
+--------------------+------------+
```

After training the model using Pyspark logistic regression model and Support Vector Machine model on pca_feature column above, and by applying model transform on the test set, one can obtain the prediction on the test data.

Finally, in order to assess the model, I defined several functions which helps calculate various metrics. There are six quantities I am going to use when evaluating the model, which are confusion matrix, F1 score, precision rate, recall rate, AUC, and overall accuracy.



```
========= Logistic Regression =========
{'TP': 8978, 'FP': 1990, 'TN': 12583, 'FN': 2425}
AUC score is 0.8254
Accuracy is 83.0%
Confusion matrix is {'TP': 8978, 'FP': 1990, 'TN': 12583, 'FN': 2425}
{'Precision': '81.86%', 'Recall': '78.73%', 'F1_score': 0.8026}

================= SVM =================
{'TP': 4103, 'FP': 114, 'TN': 14459, 'FN': 7300}
AUC score is 0.676
Accuracy is 71.46%
Confusion matrix is {'TP': 4103, 'FP': 114, 'TN': 14459, 'FN': 7300}
{'Precision': '97.3%', 'Recall': '35.98%', 'F1_score': 0.5254}
```

By the result above, we can see that logistic regression has a higher overall accuracy, while it performs the same when predicting 0s and 1s since the precision and recall rate are close. While SVM has a lower overall accuracy, it has a very high precision, which means it nearly never gives false alert of 1.

The overall performance of these model can be further improved if larger data is given. In that case, Pyspark's ability to process big dataset will be more evident.

Dataset Source:

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv