



Out[28]: 'List of airlines of the United States - Wikipedia'

In [29]: 1 soup.a

Out[29]: Jump to content

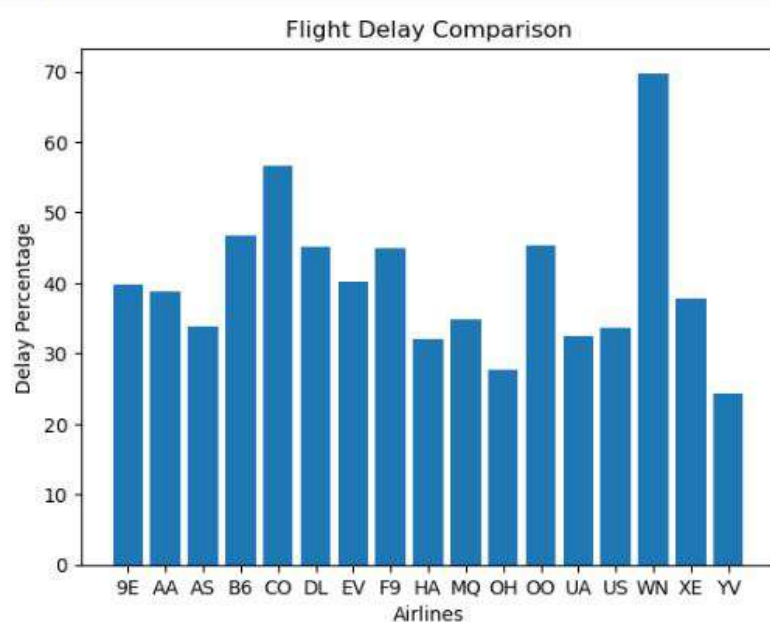
In [43]: 1 links = soup.find_all('a')

In [38]: 1 tables = soup.find_all('table')

In [44]: 1 for link in links:
2 print(link['href'])

```
#bodyContent
/wiki/Main_Page
/wiki/Wikipedia:Contents
/wiki/Portal:Current_events
/wiki/Special:Random
/wiki/Wikipedia:About
//en.wikipedia.org/wiki/Wikipedia:Contact_us
https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_source=donate&utm_medium=sidebar&utm_campaign=C13_en.wikip
edia.org&uselang=en
/wiki/Help:Contents
/wiki/Help:Introduction
/wiki/Wikipedia:Community_portal
/wiki/Special:RecentChanges
/wiki/Wikipedia:File_upload_wizard
/wiki/Main_Page
/wiki/Special:Search
/w/index.php?title=Special:CreateAccount&returnto=List+of+airlines+of+the+United+States
/w/index.php?title=Special:UserLogin&returnto=List+of+airlines+of+the+United+States
/w/index.php?title=Special:CreateAccount&returnto=List+of+airlines+of+the+United+States
```

```
In [42]: 1 # Bar Plot
2 plt.bar(delay_percentages.index, delay_percentages.values)
3 plt.xlabel('Airlines')
4 plt.ylabel('Delay Percentage')
5 plt.title('Flight Delay Comparison')
6
7
8 plt.show()
```

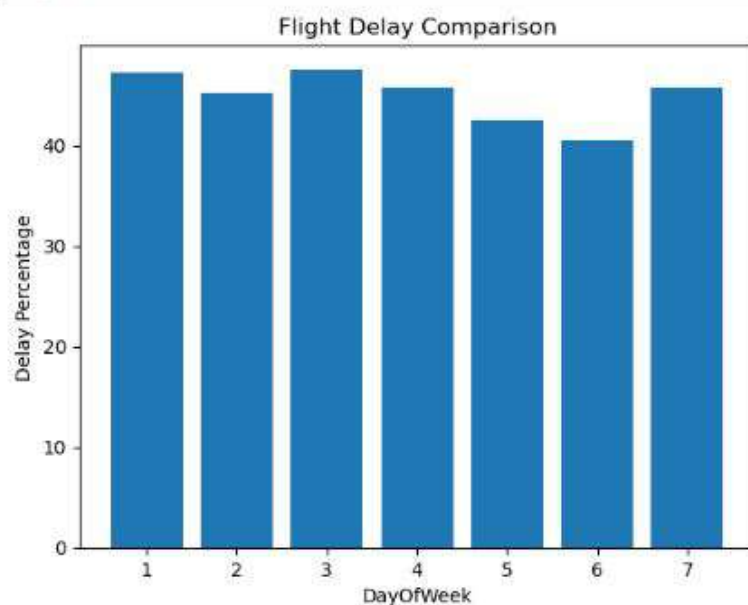


'WN' is the symbol of SouthWest Airlines and from the above graph we can see that

SouthWest airlines have high percentage of Delay time than other airlines

```
In [43]: 1 delay_week_percentages = airlines.groupby('DayOfWeek')['Delay'].mean() * 100
```

```
In [44]: 1 # Bar Plot
2 plt.bar(delay_week_percentages.index, delay_week_percentages.values)
3 plt.xlabel('DayOfWeek')
4 plt.ylabel('Delay Percentage')
5 plt.title('Flight Delay Comparison')
6
7 plt.show()
```





Long-Distance 559
 Name: Travel_Category, dtype: int64

```
In [52]: 1 pd.DataFrame(airlines.groupby(['Airline', 'Travel_Category']).size())
```

```
Out[52]:
```

0

Airline	Travel_Category	
B6	Short-Distance	20686
	Medium-Distance	0
	Long-Distance	0
AA	Short-Distance	36171
	Medium-Distance	9338
	Long-Distance	140
AS	Short-Distance	8899
	Medium-Distance	2772
	Long-Distance	0
B8	Short-Distance	14720
	Medium-Distance	3391
	Long-Distance	0
CO	Short-Distance	14803
	Medium-Distance	6340
	Long-Distance	166
DL	Short-Distance	40632
	Medium-Distance	11149
	Long-Distance	109
EV	Short-Distance	27973
	Medium-Distance	10
	Long-Distance	0
F9	Short-Distance	5932
	Medium-Distance	551
	Long-Distance	0
HA	Short-Distance	4575
	Medium-Distance	1933
	Long-Distance	0
MQ	Short-Distance	36543
	Medium-Distance	62
	Long-Distance	0
OH	Short-Distance	12506
	Medium-Distance	125
	Long-Distance	0
OO	Short-Distance	40950
	Medium-Distance	284
	Long-Distance	0
UA	Short-Distance	18157
	Medium-Distance	9327
	Long-Distance	135
US	Short-Distance	28063
	Medium-Distance	6437
	Long-Distance	0
WN	Short-Distance	66788
	Medium-Distance	0
	Long-Distance	0

```

Long-Distance 0
WN Short-Distance 88786
Medium-Distance 7311
Long-Distance 0
XB Short-Distance 31072
Medium-Distance 54
Long-Distance 0
YV Short-Distance 13667
Medium-Distance 56
Long-Distance 0

```

```

In [53]: 1 long_duration_flights = airlines[airlines['Travel Category'] == 'Long-Distance']
        2 long_duration_flights.shape

```

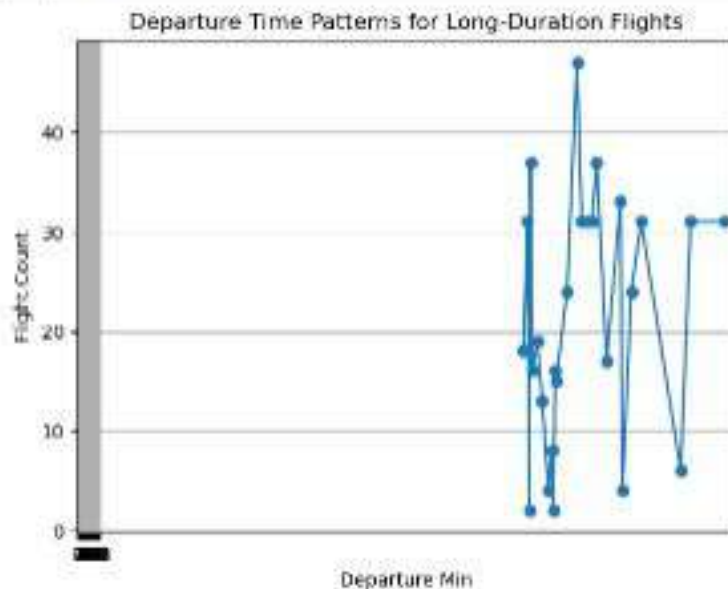
```
Out[53]: (559, 18)
```

```
In [54]: 1 departure_min_counts = long_duration_flights['length'].value_counts().sort_index()
```

```

In [55]: 1 plt.plot(departure_min_counts.index, departure_min_counts.values, marker='o')
        2 plt.xlabel('Departure Min')
        3 plt.ylabel('Flight Count')
        4 plt.title('Departure Time Patterns for Long-Duration Flights')
        5 plt.xticks(range(24)) # Set x-axis ticks for each hour
        6 plt.grid(True)
        7 plt.show()

```



Null hypothesis (H0): The duration of a flight does not affect flight delays.

Alternative hypothesis (H1): The duration of a flight affects flight delays.

```

In [56]: 1 short_delay = airlines[airlines['Travel Category'] == 'Short-Distance']['Delay']
        2 medium_delay = airlines[airlines['Travel Category'] == 'Medium-Distance']['Delay']
        3 long_delay = airlines[airlines['Travel Category'] == 'Long-Distance']['Delay']

```



Null hypothesis (H0): The duration of a flight does not affect flight delays.

Alternative hypothesis (H1): The duration of a flight affects flight delays.

```
In [56]: 1 short_delay = airlines[airlines['Travel_Category'] == 'Short-Distance']['Delay']
2 medium_delay = airlines[airlines['Travel_Category'] == 'Medium-Distance']['Delay']
3 long_delay = airlines[airlines['Travel_Category'] == 'Long-Distance']['Delay']
```

```
In [57]: 1 # Perform one-way ANOVA
2 f_stat, p_value = f_oneway(short_delay, medium_delay, long_delay)
```

```
In [58]: 1 print("F-Statistic: ", f_stat)
2 print("P-Value: ", p_value)
```

F-Statistic: 352.11904019127104
P-Value: 1.5149202534228687e-153

```
In [59]: 1 # Interpret the results
2 alpha = 0.05
3
4 if p_value < alpha:
5     print("Reject the null hypothesis.")
6     print("There is a significant difference among the delay groups.")
7 else:
8     print("Fail to reject the null hypothesis.")
9     print("There is not enough evidence to conclude a significant difference among the delay groups.")
```

Reject the null hypothesis.
There is a significant difference among the delay groups.

Null hypothesis (H0): The altitude of the airport does not affect flight delays.

Null hypothesis (H0): The altitude of the airport does not affect flight delays.

Alternative hypothesis (H1): The altitude of the airport affects flight delays.

```
In [60]: 1 t_stat, p_value = stats.ttest_ind(df['Delay'], (df['elevation_ft'].dropna()), equal_var=False)
```

```
In [61]: 1 print("F-Statistic: ", f_stat)
2 print("P-Value: ", p_value)
```

F-Statistic: 352.11904019127104
P-Value: 0.0

```
In [62]: 1 # Interpret the results
2 alpha = 0.05
3
4 if p_value < alpha:
5     print("Reject the null hypothesis.")
6     print("There is a significant difference among the delay groups.")
7 else:
8     print("Fail to reject the null hypothesis.")
9     print("There is not enough evidence to conclude a significant difference among the delay groups.")
```

Reject the null hypothesis.
There is a significant difference among the delay groups.

Null hypothesis (H0): The number of runways does not affect flight delays.

Alternative hypothesis (H1): The number of runways affects flight delays.

```
In [63]: 1 t_stat, p_value = stats.ttest_ind(df['Delay'], df['closed'], equal_var=False)
```

```
In [64]: 1 print("F-Statistic: ", f_stat)
2 print("P-Value: ", p_value)
```

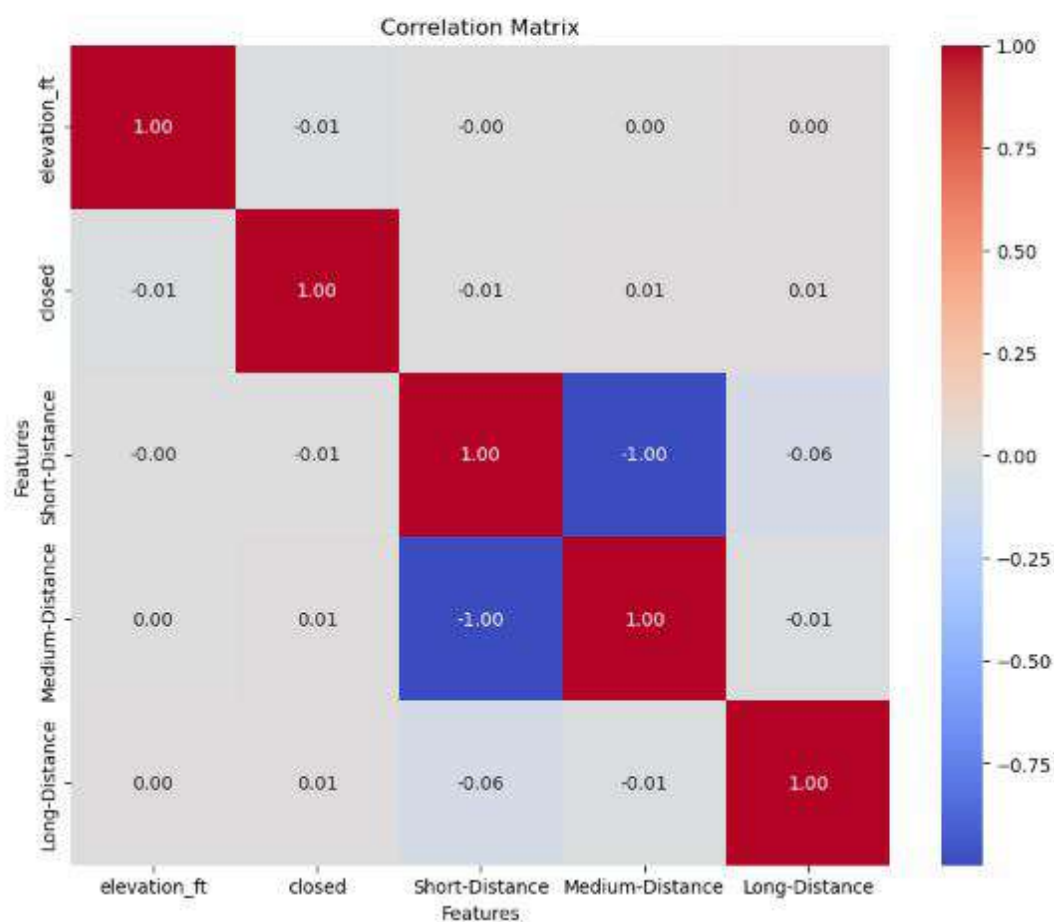
```
F-Statistic: 352.11904019127104
P-Value: 0.0
```

```
In [65]: 1 # Interpret the results
2 alpha = 0.05
3
4 if p_value < alpha:
5     print("Reject the null hypothesis.")
6     print("There is a significant difference among the delay groups.")
7 else:
8     print("Fail to reject the null hypothesis.")
9     print("There is not enough evidence to conclude a significant difference among the delay groups.")
```

Reject the null hypothesis.

There is a significant difference among the delay groups.


```
In [77]: 1 plt.figure(figsize=(10, 8))
2 sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)
3
4 plt.title('Correlation Matrix')
5 plt.xlabel('Features')
6 plt.ylabel('Features')
7
8 plt.show()
```



```
In [119]: 1 X = new_data.drop('Delay', axis=1)
          2 y = new_data['Delay']
```

```
In [137]: 1 logistic_model = SGDClassifier(loss='log_loss', random_state=42)
          2 logistic_model.fit(X, y)
```

```
Out[137]: SGDClassifier
SGDClassifier(loss='log_loss', random_state=42)
```

```
In [122]: 1 decision_tree_model = DecisionTreeClassifier(random_state=42)
          2 decision_tree_model.fit(X, y)
```

```
Out[122]: DecisionTreeClassifier
DecisionTreeClassifier(random_state=42)
```

```
In [123]: 1 n_folds = 5
          2 stratified_kfold = StratifiedKFold(n_splits=n_folds, random_state=42, shuffle=True)
          3 logistic_scores = []
          4 decision_tree_scores = []
```

```
In [128]: 1 # Perform cross-validation
          2 for train_index, val_index in stratified_kfold.split(X, y):
          3     # Split the data into training and validation sets
          4     X_train, X_val = X.iloc[train_index], X.iloc[val_index]
          5     y_train, y_val = y.iloc[train_index], y.iloc[val_index]
```

```
In [129]: 1 # Fit the logistic regression model
          2 logistic_model.fit(X_train, y_train)
          3
          4 # Predict and calculate accuracy for logistic regression model
          5 logistic_pred = logistic_model.predict(X_val)
          6 logistic_accuracy = accuracy_score(y_val, logistic_pred)
          7 logistic_scores.append(logistic_accuracy)
          8
          9 # Fit the decision tree model
         10 decision_tree_model.fit(X_train, y_train)
         11
         12 # Predict and calculate accuracy for decision tree model
         13 decision_tree_pred = decision_tree_model.predict(X_val)
         14 decision_tree_accuracy = accuracy_score(y_val, decision_tree_pred)
         15 decision_tree_scores.append(decision_tree_accuracy)
```

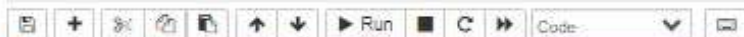
```
15 decision_tree_scores.append(decision_tree_accuracy)
```

```

1 # Define the hyperparameter grid for logistic regression
2 logistic_param_grid = {
3     'alpha': [0.001, 0.01, 0.1],
4     'max_iter': [100, 200, 300]
5 }
6
7 # Define the hyperparameter grid for decision tree
8 decision_tree_param_grid = {
9     'max_depth': [None, 5, 10],
10    'min_samples_split': [2, 5, 10]
11 }
12
13 # Create RandomizedSearchCV objects
14 logistic_search = RandomizedSearchCV(logistic_model, param_distributions=logistic_param_grid, cv=KFold(n_splits=n_folds, shuffle=True))
15 decision_tree_search = RandomizedSearchCV(decision_tree_model, param_distributions=decision_tree_param_grid, cv=KFold(n_splits=n_folds, shuffle=True))
16
17 # Fit the models with hyperparameter tuning
18 logistic_search.fit(X, y)
19 decision_tree_search.fit(X, y)
20
21 # Get the best hyperparameters for each model
22 best_logistic_params = logistic_search.best_params_
23 best_decision_tree_params = decision_tree_search.best_params_
24

```

```
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\model_selection\_search.py:306: UserWarning: The total space of parameter
s 9 is smaller than n_iter=10. Running 9 iterations. For exhaustive searches, use GridSearchCV.
warnings.warn(
```

warnings.warn()

```
In [139]: 1 # Create the SGDClassifier model
2 sgd_model = SGDClassifier(loss='log', random_state=42)
3
4 # Define the hyperparameter grid for SGDClassifier
5 sgd_param_grid = {
6     'alpha': [0.0001, 0.001, 0.01],
7     'penalty': ['l1', 'l2'],
8     'max_iter': [100, 200, 300]
9 }
10
11 # Create the RandomizedSearchCV object for SGDClassifier
12 sgd_search = RandomizedSearchCV(sgd_model, param_distributions=sgd_param_grid, cv=KFold(n_splits=n_folds, shuffle=True), ran
13
14 # Fit the model with hyperparameter tuning
15 sgd_search.fit(X, y)
16
```

```
warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:173: FutureWarning: The loss 'log' w
as deprecated in v1.1 and will be removed in version 1.3. Use 'loss='log_loss' which is equivalent.
    warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
    warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:173: FutureWarning: The loss 'log' w
as deprecated in v1.1 and will be removed in version 1.3. Use 'loss='log_loss' which is equivalent.
    warnings.warn(
C:\Users\naray\anaconda3\lib\site-packages\sklearn\linear_model\_stochastic_gradient.py:705: ConvergenceWarning: Maximum num
ber of iteration reached before convergence. Consider increasing max_iter to improve the fit.
    warnings.warn(
```

```
Out[139]: RandomizedSearchCV
  estimator: SGDClassifier
    SGDClassifier
```

```
In [152]: 1 xgb_classifier = XGBClassifier()
```

```
In [155]: 1 n_folds = 5
2 stratified_kfold = StratifiedKFold(n_splits=n_folds, shuffle=True, random_state=42)
3
4 # Perform cross-validation using the XGBoost classifier
5 xgb_scores = cross_val_score(xgb_classifier, X, y, cv=stratified_kfold)
```

SGDClassifier

```
In [152]: 1 xgb_classifier = XGBClassifier()
```

```
In [155]: 1 n_folds = 5
2 stratified_kfold = StratifiedKFold(n_splits=n_folds, shuffle=True, random_state=42)
3
4 # Perform cross-validation using the XGBoost classifier
5 xgb_scores = cross_val_score(xgb_classifier, X, y, cv=stratified_kfold)
```

```
In [147]: 1 best_logistic_params
```

```
Out[147]: {'max_iter': 200, 'alpha': 0.01}
```

```
In [148]: 1 best_decision_tree_params
```

```
Out[148]: {'min_samples_split': 2, 'max_depth': None}
```

```
In [144]: 1 best_sgd_params = sgd_search.best_params_
2 best_sgd_params
```

```
Out[144]: {'penalty': 'l2', 'max_iter': 100, 'alpha': 0.01}
```

```
In [150]: 1 logistic_mean_accuracy = sum(logistic_scores) / len(logistic_scores)
2 decision_tree_mean_accuracy = sum(decision_tree_scores) / len(decision_tree_scores)
3
4 print("Logistic Regression Mean Accuracy:", logistic_mean_accuracy)
5 print("Decision Tree Mean Accuracy:", decision_tree_mean_accuracy)
```

```
Logistic Regression Mean Accuracy: 0.5807574327512978
Decision Tree Mean Accuracy: 0.6493629070316187
```

```
In [156]: 1 print("XGBoost Classifier Mean Accuracy:", xgb_scores.mean())
```

```
XGBoost Classifier Mean Accuracy: 0.6683181451539499
```

```
In [ ]: 1
```

US Airlines Analysis



Total Count by Category

Airline	Travel Category		
	Long-Di..	Mediu..	Short-D..
9E			1,684
AA	8	731	2,935
AS		246	710
B6		242	1,098
CO	6	483	1,131
DL	3	903	4,099
EV		1	2,317
F9		41	445
HA		81	363
MQ		6	3,013
OH		4	1,053
OO		19	4,145
UA	4	815	1,544
US		512	2,455
WN		605	7,213
XE		2	2,451
YV		5	1,011

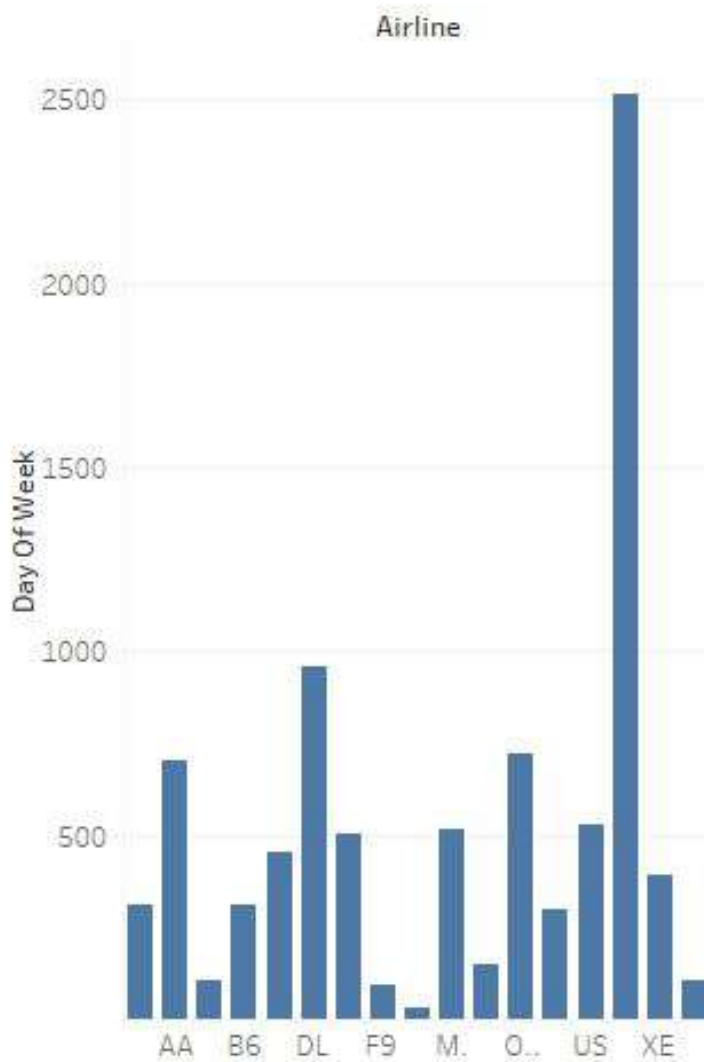
Number of Delays by Category

Airline	Travel Category		
	Long-Di..	Mediu..	Short-D..
9E			595
AA	4	306	904
AS		84	171
B6		110	494
CO	4	312	521
DL	2	386	1,409
EV		1	815
F9		15	156
HA		29	54
MQ		2	908
OH		1	250
OO		4	1,542
UA	0	219	294
US		197	740
WN		450	4,250
XE		2	689
YV		1	165

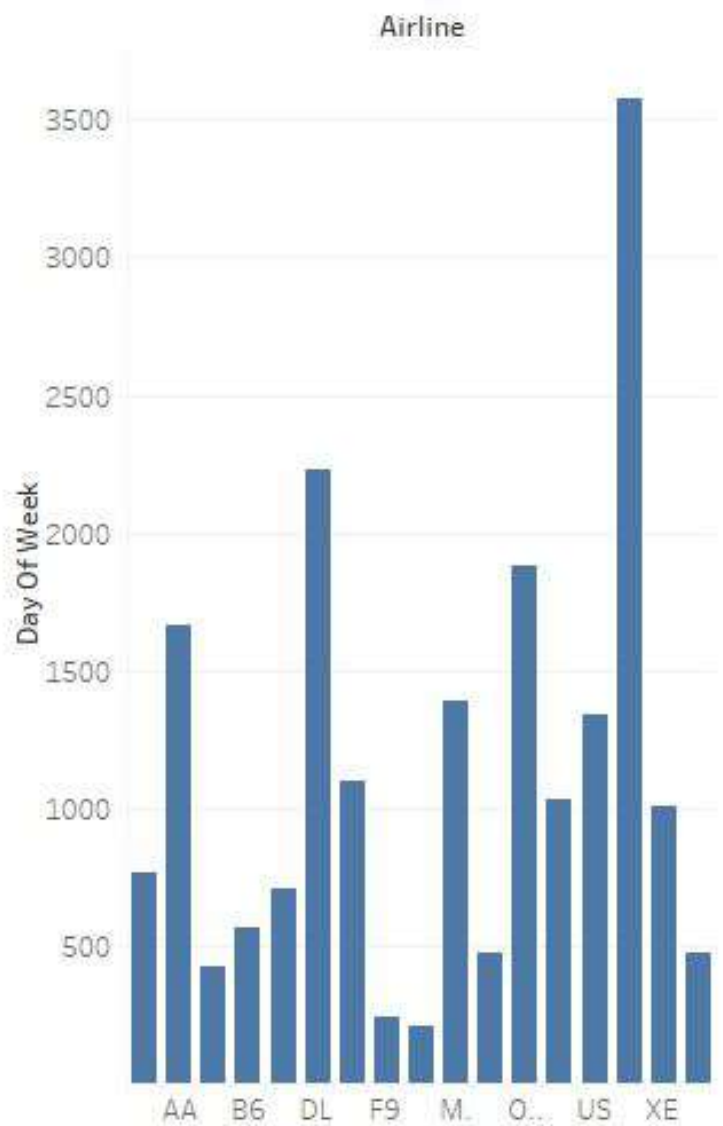
US Airlines Analysis



Number of Delays by Day of Week



Total Count by Day of Week



< ● ● ● ● >

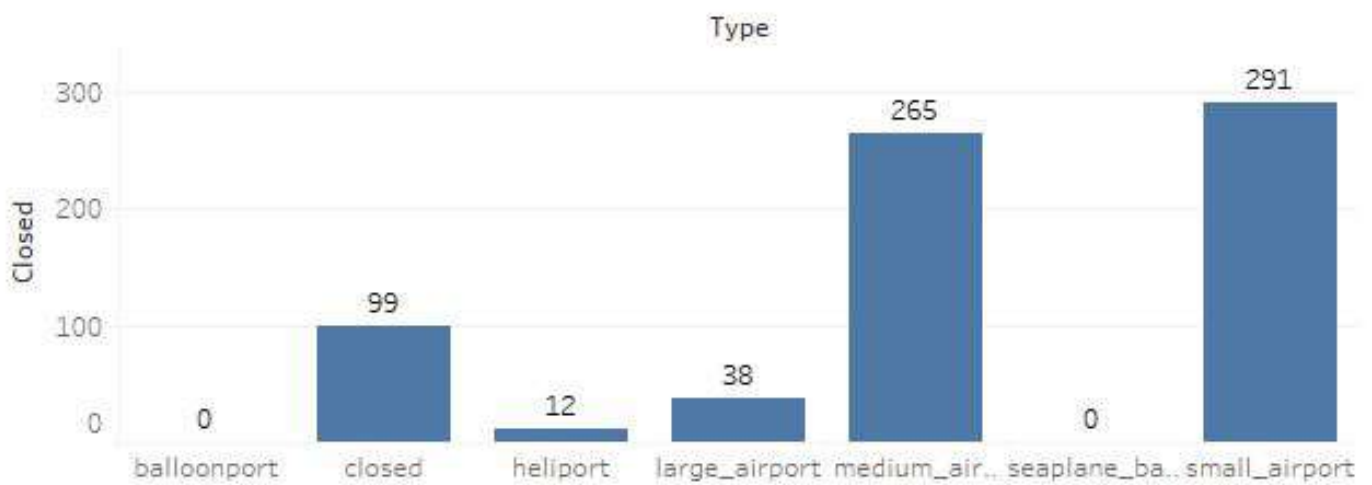
	Airline								
Type	9E	AA	AS	B6	CO	DL	EV	F9	HA
balloonport	2	1		1	1	2			1
closed	101	241	71	78	109	315	136	32	48
heliport	274	634	151	211	276	760	405	71	61
large_airpo..	40	65	17	27	34	83	60	14	16
medium_ai..	258	555	149	181	221	682	342	57	63
seaplane_b..	30	46	8	23	22	112	26	8	13
small_airpo..	979	2,132	560	819	957	3,051	1,349	304	242

	Airline								
Type	9E	AA	AS	B6	CO	DL	EV	F9	HA
balloonport	0	0		1	1	2			^
closed	34	103	16	43	57	137	48	7	
heliport	97	247	35	90	151	280	151	22	1
large_airpo..	14	20	10	8	22	35	10	4	
medium_ai..	85	171	50	80	131	240	149	30	1
seaplane_b..	18	17	1	14	15	44	5	3	
small_airpo..	347	656	143	368	460	1,059	453	105	4

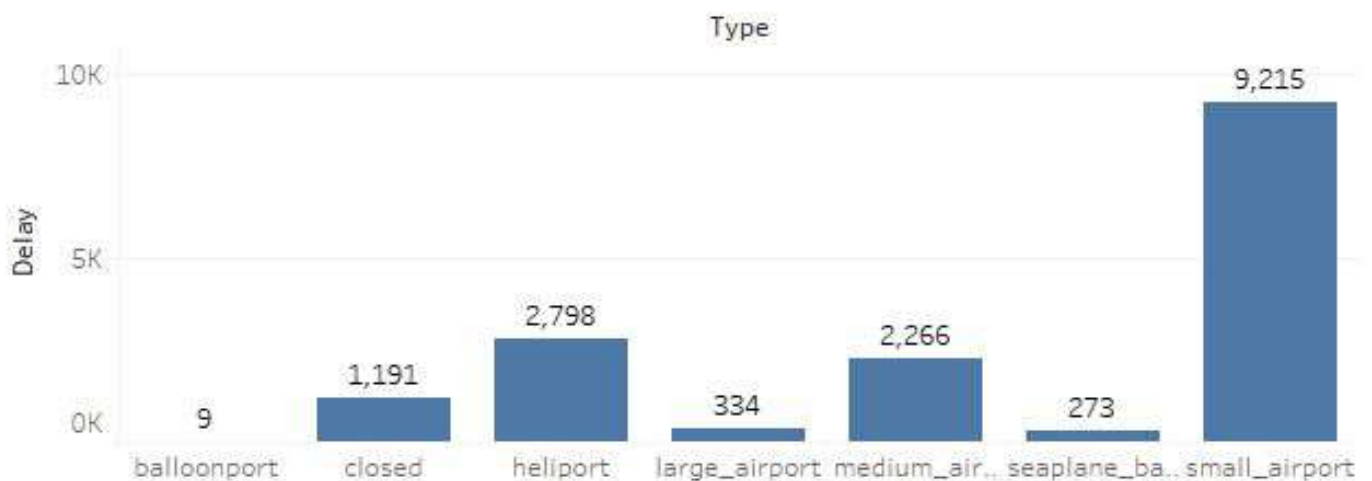
US Airlines Analysis



Number of Runways closed with respect to types of Airports



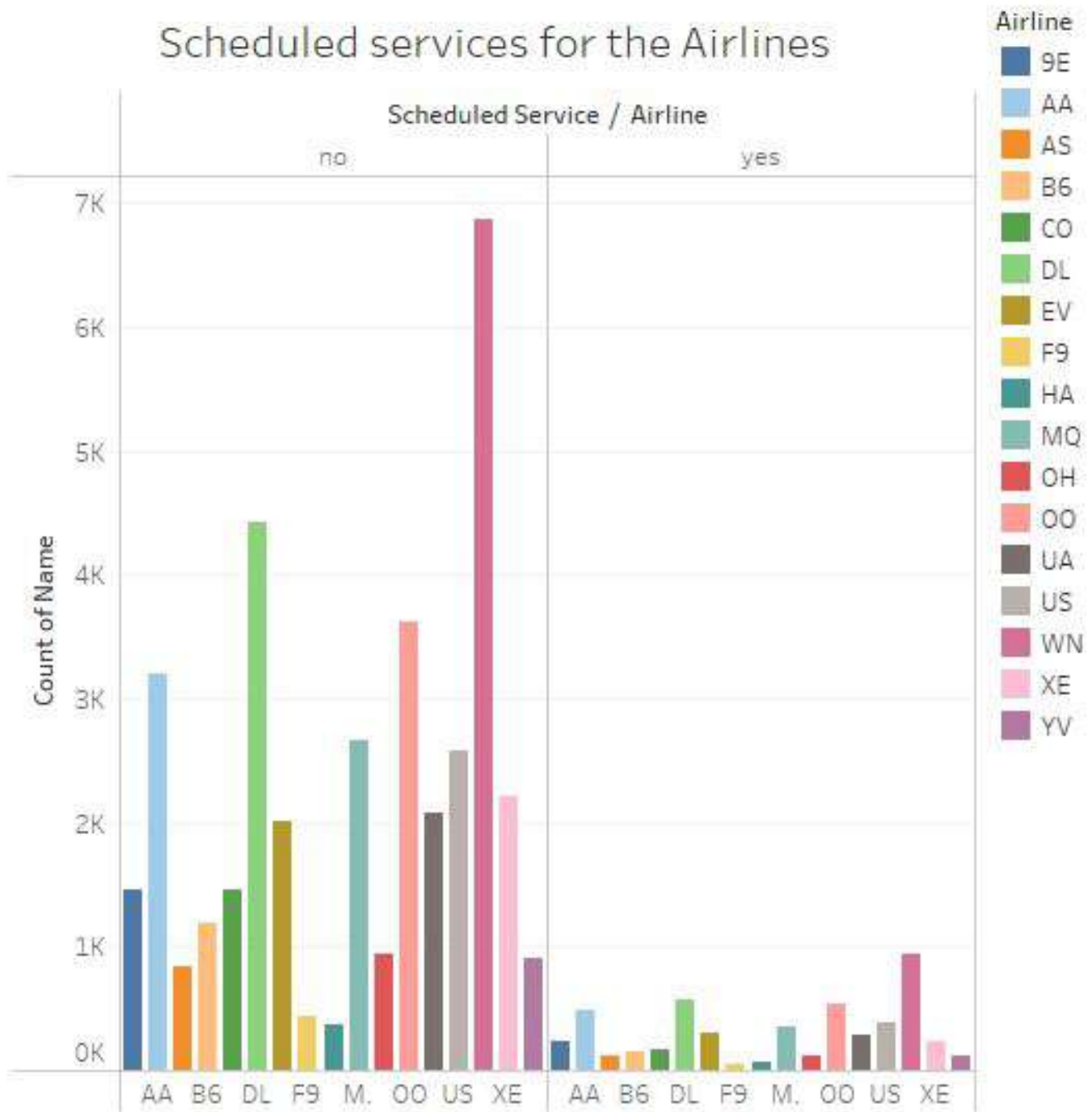
Flight delays with respect to types of Airports



US Airlines Analysis



Scheduled services for the Airlines



F	G	H	I
	On-time Performance		
	Airline	<input type="button" value="▼"/> Sum of Length	
	9E	2.91%	
	AA	11.28%	
	AS	3.01%	
	B6	4.06%	
	CO	5.52%	
	DL	13.96%	
	EV	3.81%	
	F9	1.32%	
	HA	0.76%	
	MQ	5.63%	
	OH	2.06%	
	OO	7.05%	
	UA	8.07%	
	US	7.63%	
	WN	16.49%	
	XE	4.79%	
	YV	1.64%	
	Grand Total	100.00%	

Percentage of Delayed flights with respective to week

Count of id_x	Delay		
DayOfWeek	On-Time	Delayed	Grand Total
Sunday	1.34%	1.42%	2.77%
Monday	0.97%	0.91%	1.89%
Tuesday	24.50%	20.43%	44.93%
Wednesday	23.68%	9.95%	33.62%
Thursday	9.31%	2.88%	12.19%
Friday	0.74%	0.86%	1.60%
Saturday	1.51%	1.50%	3.00%
Grand Total	62.05%	37.95%	100.00%

Count of Id_x		Delay		
AirportFrom	AirportTo	On Time	Delayed	Grand Total
PHX		754	410	1164
CLT		13	9	22
DFW		39	5	44
SEA		17	19	36
MSP		20	11	31
DTW		14	9	23
ORD		19	9	28
ATL		18	3	21
PDX		14	13	27
JFK		8	5	13
IAH		17	5	22
SJC		30	16	46
HNL		6	1	7
MCO		6	2	8
OGG		3	3	6
LAX		31	17	48
KOA		1	1	2
SFO		19	12	31
MIA		2	1	3
LAD		2	2	4
SMS		12	8	20
PHL		14	6	20
LHR		1	1	2
DEN		36	12	48
MEM		7	1	8
CVG		1	4	5
YUM		10	10	20
MKE		3	1	4
LAS		41	19	60
ANC		1	1	2
BOS		3	11	14
LGB		12	3	15
FIL		3	2	5
EWB		9	8	17
DCA		3	1	4
RDU		3	2	5
MCI		6	7	13
SAN		17	19	36
ONT		14	13	27
OAK		15	6	21

