Project - Big Data Analytics and Visualization

<u>Domain:</u> Flight Travel Agency Service.

<u>Task:</u> Perform whether flight will be delayed or not.

Problem Statement

Business Requirement

Sky Travel (ST) provides concierge services for business travelers. In an increasingly crowded market, they are always looking for ways to differentiate themselves and provide added value to their corporate customers.

ST is investigating ways that they can capitalize on their existing data assets to provide new insights that provide them a strategic advantage against their competition. In planning their product, they heard much fanfare about machine learning and came up with the idea of using predictive analytics to help customers best select their travels based on the likelihood of a delay. When reviewing their customer transaction histories, they discovered that their most premium customers often book their travel within 7 days of departure. In speaking with customer service, they learned that these customers often ask questions like, "I don't have to be there until Tuesday, so is it better for me to fly out on Sunday or Monday?"

While there are many factors that customer service uses to tailor their guidance to the customer (such as cost and travel duration), ST believes an innovative solution might come in the form of giving the customer an assessment of the risk of encountering flight delays. For low-risk flights, the customer may choose to book flight, giving them more precious time at home and less on the road spent arriving too early to a destination. ST is interested in applying data science to the problem to discover if the weather forecast coupled with their historical flight delay data could be used to provide a meaningful input into the customer's decision-making process.

Business Requirement

1. Generate detailed reports that can be visualized in a dashboard. The detailed reports can include tables that are a subset or findings from the dataset.

2. Build an automized solution using machine learning models whether flight will be delayed or not.

3. Create scheduled pipeline for new data arriving.

Dataset

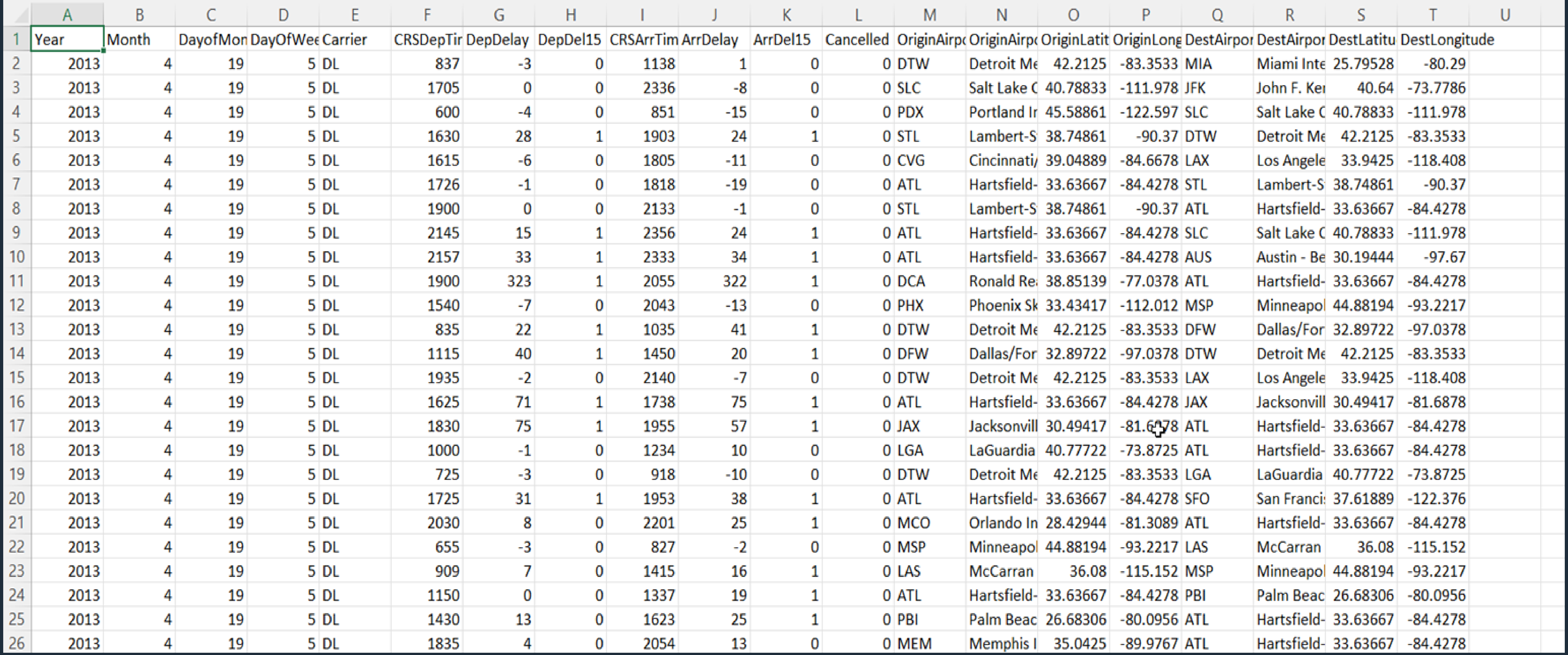
1. 'FlightDelaysWithAirportCodes.csv'
It contains information about the historical flight delays.

2. 'FlightWeatherWithAirportCode.csv'
It contains all the information of the weather corresponding to flight delays.

3. 'AirportCodeLocationLookupClean.csv'

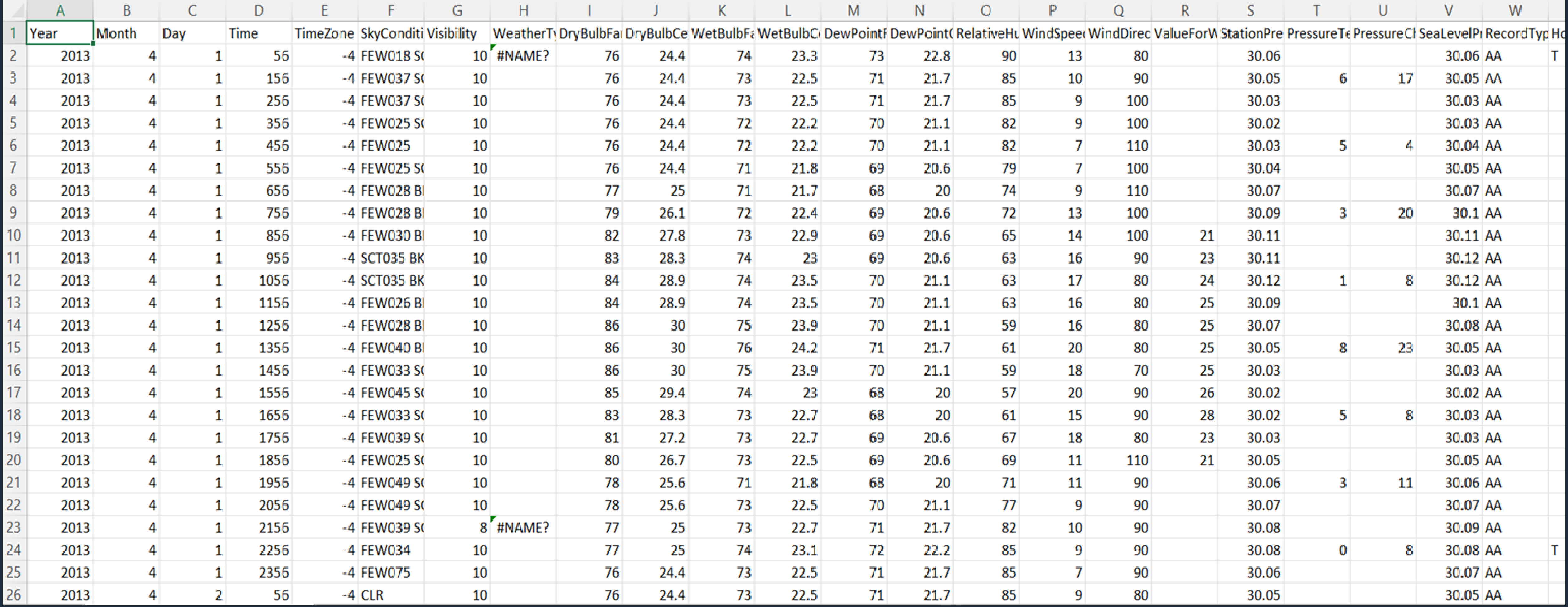
It contains all the information regarding Airport.

Snapshots of Dataset: 'FlightDelaysWithAirportCodes.csv'



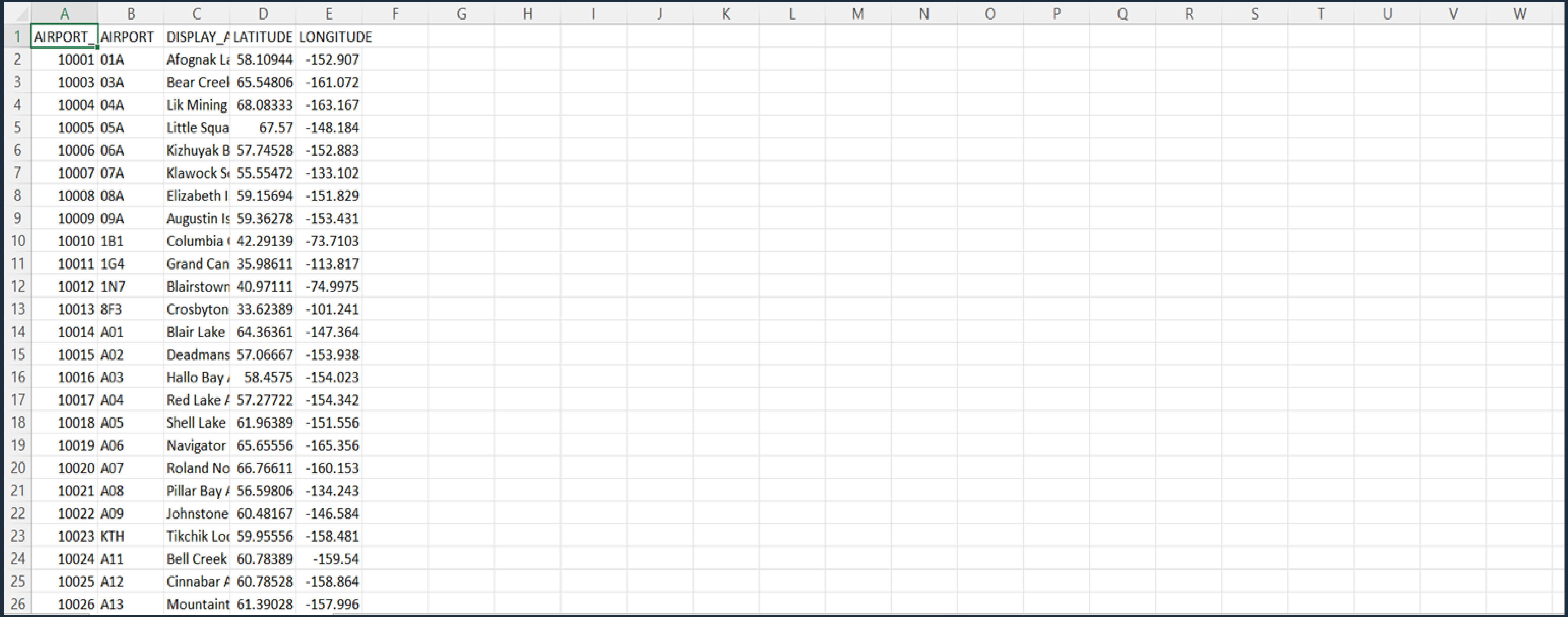
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Year	Month	DayofMon	DayOfWee	Carrier	CRSDepTir	DepDelay	DepDel15	CRSArrTim	ArrDelay	ArrDel15	Cancelled	OriginAirpc	OriginAirpc	OriginLatit	OriginLong	DestAirpor	DestAirpor	DestLatitu	DestLongitude	
2	2013	4	19	5	DL	837	-3	0	1138	1	0	0	DTW	Detroit Me	42.2125	-83.3533	MIA	Miami Inte	25.79528	-80.29	
3	2013	4	19	5	DL	1705	0	0	2336	-8	0	0	SLC	Salt Lake C	40.78833	-111.978	JFK	John F. Ker	40.64	-73.7786	
4	2013	4	19	5	DL	600	-4	0	851	-15	0	0	PDX	Portland Ir	45.58861	-122.597	SLC	Salt Lake C	40.78833	-111.978	
5	2013	4	19	5	DL	1630	28	1	1903	24	1	0	STL	Lambert-S	38.74861	-90.37	DTW	Detroit Me	42.2125	-83.3533	
6	2013	4	19	5	DL	1615	-6	0	1805	-11	0	0	CVG	Cincinnati/	39.04889	-84.6678	LAX	Los Angele	33.9425	-118.408	
7	2013	4	19	5	DL	1726	-1	0	1818	-19	0	0	ATL	Hartsfield-	33.63667	-84.4278	STL	Lambert-S	38.74861	-90.37	
8	2013	4	19	5	DL	1900	0	0	2133	-1	0	0	STL	Lambert-S	38.74861	-90.37	ATL	Hartsfield-	33.63667	-84.4278	
9	2013	4	19	5	DL	2145	15	1	2356	24	1	0	ATL	Hartsfield-	33.63667	-84.4278	SLC	Salt Lake C	40.78833	-111.978	
10	2013	4	19	5	DL	2157	33	1	2333	34	1	0	ATL	Hartsfield-	33.63667	-84.4278	AUS	Austin - Be	30.19444	-97.67	
11	2013	4	19	5	DL	1900	323	1	2055	322	1	0	DCA	Ronald Rea	38.85139	-77.0378	ATL	Hartsfield-	33.63667	-84.4278	
12	2013	4	19	5	DL	1540	-7	0	2043	-13	0	0	PHX	Phoenix Sk	33.43417	-112.012	MSP	Minneapol	44.88194	-93.2217	
13	2013	4	19	5	DL	835	22	1	1035	41	1	0	DTW	Detroit Me	42.2125	-83.3533	DFW	Dallas/For	32.89722	-97.0378	
14	2013	4	19	5	DL	1115	40	1	1450	20	1	0	DFW	Dallas/For	32.89722	-97.0378	DTW	Detroit Me	42.2125	-83.3533	
15	2013	4	19	5	DL	1935	-2	0	2140	-7	0	0	DTW	Detroit Me	42.2125	-83.3533	LAX	Los Angele	33.9425	-118.408	
16	2013	4	19	5	DL	1625	71	1	1738	75	1	0	ATL	Hartsfield-	33.63667	-84.4278	JAX	Jacksonvill	30.49417	-81.6878	
17	2013	4	19	5	DL	1830	75	1	1955	57	1	0	JAX	Jacksonvill	30.49417	-81.6878	ATL	Hartsfield-	33.63667	-84.4278	
18	2013	4	19	5	DL	1000	-1	0	1234	10	0	0	LGA	LaGuardia	40.77722	-73.8725	ATL	Hartsfield-	33.63667	-84.4278	
19	2013	4	19	5	DL	725	-3	0	918	-10	0	0	DTW	Detroit Me	42.2125	-83.3533	LGA	LaGuardia	40.77722	-73.8725	
20	2013	4	19	5	DL	1725	31	1	1953	38	1	0	ATL	Hartsfield-	33.63667	-84.4278	SFO	San Franci	37.61889	-122.376	
21	2013	4	19	5	DL	2030	8	0	2201	25	1	0	MCO	Orlando In	28.42944	-81.3089	ATL	Hartsfield-	33.63667	-84.4278	
22	2013	4	19	5	DL	655	-3	0	827	-2	0	0	MSP	Minneapol	44.88194	-93.2217	LAS	McCarran	36.08	-115.152	
23	2013	4	19	5	DL	909	7	0	1415	16	1	0	LAS	McCarran	36.08	-115.152	MSP	Minneapol	44.88194	-93.2217	
24	2013	4	19	5	DL	1150	0	0	1337	19	1	0	ATL	Hartsfield-	33.63667	-84.4278	PBI	Palm Beac	26.68306	-80.0956	
25	2013	4	19	5	DL	1430	13	0	1623	25	1	0	PBI	Palm Beac	26.68306	-80.0956	ATL	Hartsfield-	33.63667	-84.4278	
26	2013	4	19	5	DL	1835	4	0	2054	13	0	0	MEM	Memphis I	35.0425	-89.9767	ATL	Hartsfield-	33.63667	-84.4278	

Snapshots of Dataset: 'FlightWeatherWithAirportCode.csv'



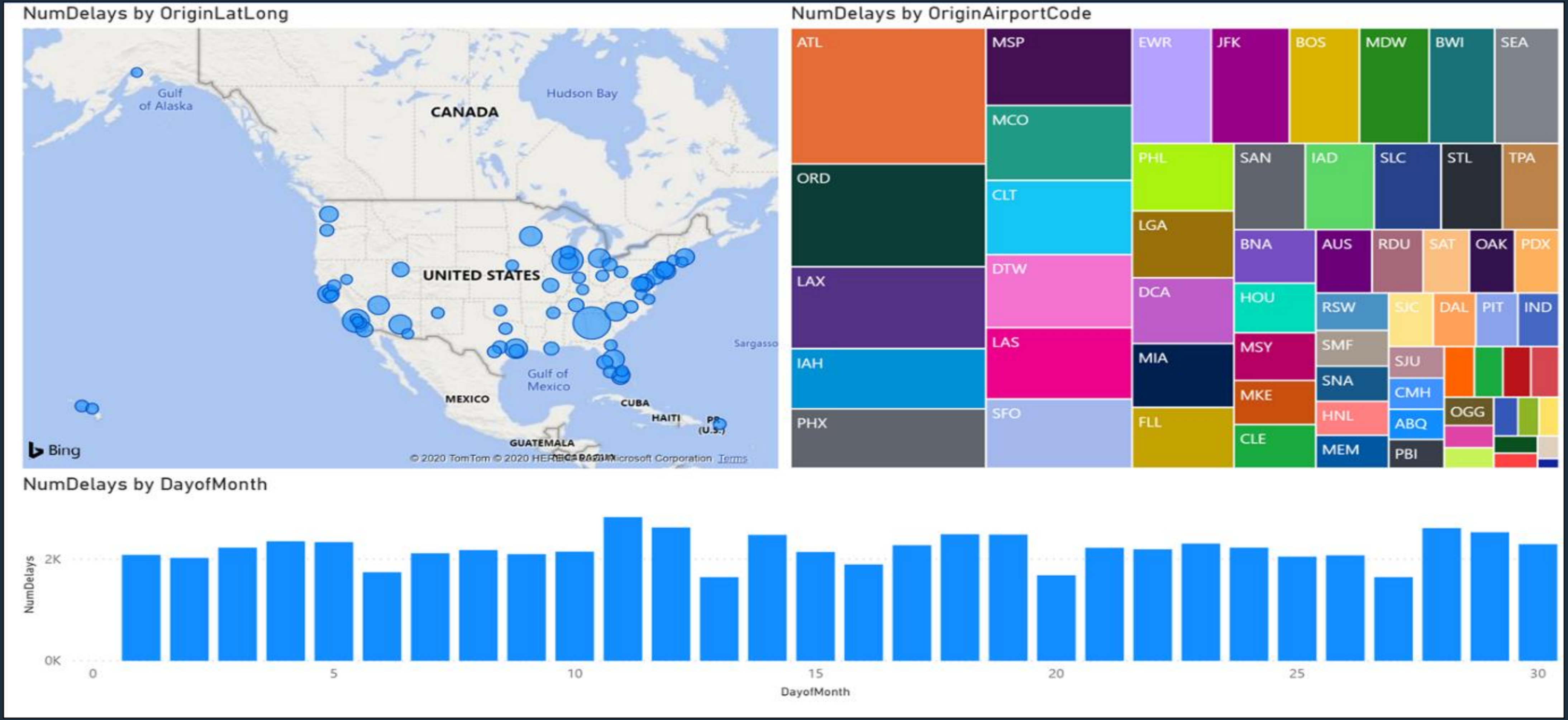
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	Year	Month	Day	Time	TimeZone	SkyConditi	Visibility	WeatherTy	DryBulbFa	DryBulbCe	WetBulbFa	WetBulbCe	DewPointF	DewPointC	RelativeHu	WindSpeed	WindDirec	ValueForW	StationPre	PressureTe	PressureCl	SeaLevelPr	RecordTyp	Hc
2	2013	4	1	56	-4	FEW018 SC	10	#NAME?	76	24.4	74	23.3	73	22.8	90	13	80		30.06			30.06	AA	T
3	2013	4	1	156	-4	FEW037 SC	10		76	24.4	73	22.5	71	21.7	85	10	90		30.05	6	17	30.05	AA	
4	2013	4	1	256	-4	FEW037 SC	10		76	24.4	73	22.5	71	21.7	85	9	100		30.03			30.03	AA	
5	2013	4	1	356	-4	FEW025 SC	10		76	24.4	72	22.2	70	21.1	82	9	100		30.02			30.03	AA	
6	2013	4	1	456	-4	FEW025	10		76	24.4	72	22.2	70	21.1	82	7	110		30.03	5	4	30.04	AA	
7	2013	4	1	556	-4	FEW025 SC	10		76	24.4	71	21.8	69	20.6	79	7	100		30.04			30.05	AA	
8	2013	4	1	656	-4	FEW028 BK	10		77	25	71	21.7	68	20	74	9	110		30.07			30.07	AA	
9	2013	4	1	756	-4	FEW028 BK	10		79	26.1	72	22.4	69	20.6	72	13	100		30.09	3	20	30.1	AA	
10	2013	4	1	856	-4	FEW030 BK	10		82	27.8	73	22.9	69	20.6	65	14	100	21	30.11			30.11	AA	
11	2013	4	1	956	-4	SCT035 BK	10		83	28.3	74	23	69	20.6	63	16	90	23	30.11			30.12	AA	
12	2013	4	1	1056	-4	SCT035 BK	10		84	28.9	74	23.5	70	21.1	63	17	80	24	30.12	1	8	30.12	AA	
13	2013	4	1	1156	-4	FEW026 BK	10		84	28.9	74	23.5	70	21.1	63	16	80	25	30.09			30.1	AA	
14	2013	4	1	1256	-4	FEW028 BK	10		86	30	75	23.9	70	21.1	59	16	80	25	30.07			30.08	AA	
15	2013	4	1	1356	-4	FEW040 BK	10		86	30	76	24.2	71	21.7	61	20	80	25	30.05	8	23	30.05	AA	
16	2013	4	1	1456	-4	FEW033 SC	10		86	30	75	23.9	70	21.1	59	18	70	25	30.03			30.03	AA	
17	2013	4	1	1556	-4	FEW045 SC	10		85	29.4	74	23	68	20	57	20	90	26	30.02			30.02	AA	
18	2013	4	1	1656	-4	FEW033 SC	10		83	28.3	73	22.7	68	20	61	15	90	28	30.02	5	8	30.03	AA	
19	2013	4	1	1756	-4	FEW039 SC	10		81	27.2	73	22.7	69	20.6	67	18	80	23	30.03			30.03	AA	
20	2013	4	1	1856	-4	FEW025 SC	10		80	26.7	73	22.5	69	20.6	69	11	110	21	30.05			30.05	AA	
21	2013	4	1	1956	-4	FEW049 SC	10		78	25.6	71	21.8	68	20	71	11	90		30.06	3	11	30.06	AA	
22	2013	4	1	2056	-4	FEW049 SC	10		78	25.6	73	22.5	70	21.1	77	9	90		30.07			30.07	AA	
23	2013	4	1	2156	-4	FEW039 SC	8	#NAME?	77	25	73	22.7	71	21.7	82	10	90		30.08			30.09	AA	
24	2013	4	1	2256	-4	FEW034	10		77	25	74	23.1	72	22.2	85	9	90		30.08	0	8	30.08	AA	T
25	2013	4	1	2356	-4	FEW075	10		76	24.4	73	22.5	71	21.7	85	7	90		30.06			30.07	AA	
26	2013	4	2	56	-4	CLR	10		76	24.4	73	22.5	71	21.7	85	9	80		30.05			30.05	AA	

Snapshots of Dataset: 'AirportCodeLocationLookupClean.csv'



	A	B	C	D	E
1	AIRPORT_	AIRPORT	DISPLAY_A	LATITUDE	LONGITUDE
2	10001	01A	Afognak La	58.10944	-152.907
3	10003	03A	Bear Creel	65.54806	-161.072
4	10004	04A	Lik Mining	68.08333	-163.167
5	10005	05A	Little Squa	67.57	-148.184
6	10006	06A	Kizhuyak B	57.74528	-152.883
7	10007	07A	Klawock S	55.55472	-133.102
8	10008	08A	Elizabeth I	59.15694	-151.829
9	10009	09A	Augustin Is	59.36278	-153.431
10	10010	1B1	Columbia (	42.29139	-73.7103
11	10011	1G4	Grand Can	35.98611	-113.817
12	10012	1N7	Blairstown	40.97111	-74.9975
13	10013	8F3	Crosbyton	33.62389	-101.241
14	10014	A01	Blair Lake	64.36361	-147.364
15	10015	A02	Deadmans	57.06667	-153.938
16	10016	A03	Hallo Bay A	58.4575	-154.023
17	10017	A04	Red Lake A	57.27722	-154.342
18	10018	A05	Shell Lake	61.96389	-151.556
19	10019	A06	Navigator	65.65556	-165.356
20	10020	A07	Roland No	66.76611	-160.153
21	10021	A08	Pillar Bay A	56.59806	-134.243
22	10022	A09	Johnstone	60.48167	-146.584
23	10023	KTH	Tikchik Loc	59.95556	-158.481
24	10024	A11	Bell Creek	60.78389	-159.54
25	10025	A12	Cinnabar A	60.78528	-158.864
26	10026	A13	Mountaint	61.39028	-157.996

Sample visualizations



NumDelays by OriginLatLong

NumDelays by OriginAirportCode

NumDelays by DayofMonth

Step 1

- Understand the business problem.

- Understanding the columns of the dataset and identify the target variable.

- Formulate the objectives to accomplish the business problem.

- Prepare a solution plan for solving the problem.

- **Submissions:** A ppt comprising objectives (for both business requirements), and a high-level block diagram representing the solution. Selected candidates will be presenting the ppt from 1:00 PM

Step 2

- Identify the Azure services required to implement the solution.
- Develop the architecture to solve the problem using services of Azure.
- Make a list of operations/actions to be performed to accomplish the solution.

Thank You