



## **Data Collection and Preprocessing Phase**

Date	15 July 2024
Team ID	739791
Project Title	Flight Delay Prediction using Machine Learning.
Maximum Marks	6 Marks

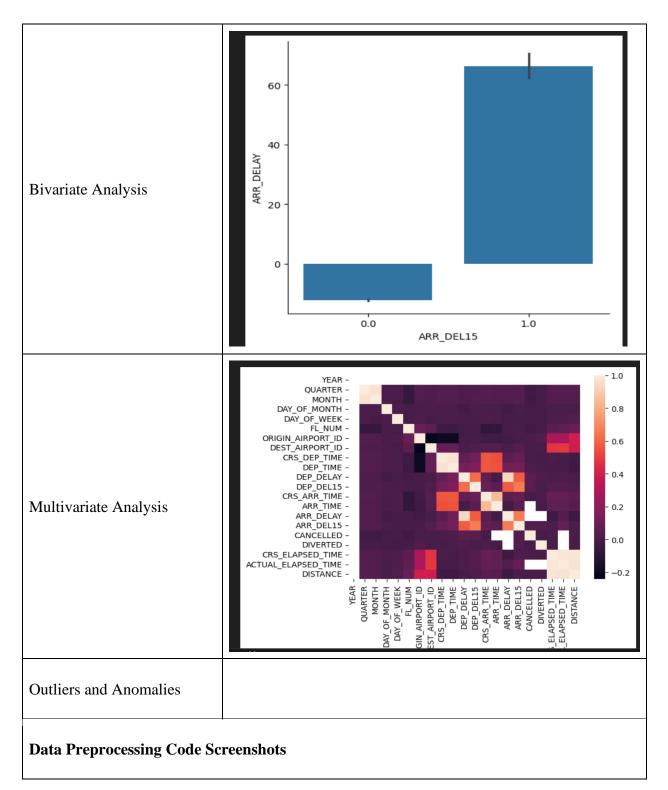
## **Data Exploration and Preprocessing Template**

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Descr	iption								
	YEA	R QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEER	C FL_NUM	ORIGIN_AIRP	PORT_ID DEST_AIRPO	RT_ID CRS_DEP_TIME	DEP_TIME
	count 11231.	0 11231.000000	11231.000000	11231.000000	11231.00000	11231.000000	11231	.000000 11231.0	00000 11231.000000	11124.000000
	mean 2016.	0 2.544475	6.628973	15.790758	3.960199	1334.325617	12334	1.516695 12302.2	74508 1320.798326	1327.189410
	std 0.	0 1.090701	3.354678	8.782056	1.99525	7 811.875227	1595	5.026510 1601.9	88550 490.737845	500.306462
	min 2016.		1.000000	1.000000				7.000000 10397.0		1.000000
	25% 2016		4.000000	8.000000		624.000000	10397	7.000000 10397.0	00000 905.000000	905.000000
	50% 2016.		7.000000	16.000000				3.000000 12478.0		1324.000000
	75% 2016.	0 3.000000	9.000000	23.000000	6.00000	2032.000000	13487	7.000000 13487.0	00000 1735.000000	1739.000000
	max 2016.	0 4.000000	12.000000	31.000000	7.00000	2853.000000	14747	7.000000 14747.0	00000 2359.000000	2400.000000
Data Overview	8 rows × 21 colu									
		CRS_ARR_TIME	ARR_TIME			CANCELLED			CTUAL_ELAPSED_TIME	
	11124.000000	11231.000000			11045.000000 1		1231.000000	11231.000000		11231.000000
	0.142844	1537.312795 502.512494	1523.978499 512.536041	-2.573123 39.232521	0.124672 0.330361	0.010150 0.100241	0.006589	190.652124 78.386317	179.661233 77.940399	1161.03196 643.68337
	0.000000	2.000000	1.000000	-67.000000	0.000000	0.000000	0.000000	93.000000	75.000000	509.000000
	0.000000	1130.000000	1135.000000	-19.000000	0.000000	0.000000	0.000000	127.000000	117.000000	594.000000
	0.000000	1559.000000	1547.000000	-10.000000	0.000000	0.000000	0.000000	159.000000	149.000000	907.00000
	0.000000	1952.000000	1945.000000	1.000000	0.000000	0.000000	0.000000	255.000000	236.000000	1927.00000
	1.000000	2359.000000	2400.000000	615.000000	1.000000	1.000000	1.000000	397.000000	428.000000	2422.00000
Univariate Analysis	p1	ights['t.show(	YEAR']	b.pyplovalue_c		t.plot(k	kind='p	ie',autopo	t='%.0f')	











	C2.4 1											
	flight	its=pd.read its		icuaca.csv )								
	,	YEAR QUA	RTER MON	TH DAY_OF_MC	ONTH DAY_C	F_WEEK UNI	QUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_	ID ORIGIN	. DEP_DEL15
		2016						N836DN	1399	103		. 0.0
		2016 2016					DL DL	N964DN N813DN	1476 1597	114		. 0.0 . 0.0
	_	2016						N587NW		147		0.0
	4	2016						N836DN	1823	147	47 SEA	. 0.0
		2016						 N940DL		114	DTW	. 0.0
												. 1.0
		2016 2016					DL DL	N583NW N554NW	1823 1901	114		. 0.0 . 0.0
	11220	2016	4	12	30	,	DI	NBASDN	2005	103	97 ATI	0.0
oading Data												
8	S ARR	TIME A	RR TIME	ARR DELAY	ARR_DEL15	CANCELLE	ED DIVERTE	D CRS E	LAPSED TI	IME ACTUAL E	ELAPSED_TIME	E DISTANCE
		2143	2102.0	-41.0	0.0					338	295.0	
			1439.0									
		1215	1142.0	-33.0	0.0					335	300.0	
		1335 607	1345.0 615.0	10.0 8.0	0.0			0		196 247	205.0 259.0	
					0.0						255.	
		1223	1148.0	-35.0	0.0					138	105.0	594
		2046	2100.0		0.0					201	181.0	
		2210 1806	2154.0 1801.0	-16.0 -5.0	0.0					311 336	295.0 332.0	
		925	913.0	-12.0	0.0			0		120	110.0	
		ights=fl:	ights.fil	lna({'ARR_DE								
	f1:	ights=fl: ights=fl: ights.ile	ights.fil	lna({'ARR_DE lna({'dep_de								ī
	f1: f1: ✓ 0.0	ights=fl: ights=fl: ights.ilo os FL_NUM	ights.fil	lna({'ARR_DE lna({'dep_de	:115':0}) DNTH DAY	/_OF_WEEK			S_ARR_TIM			
	f1: f1:	ights=fl; ights=fl; ights.ild os FL_NUM 2834	ights.fil ights.fil oc[177:18	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]		<b>/_OF_WEEK</b> 6 6	MSP	SEA	- 8	52 0.	0 1	.0
	f1: f1: ✓ 0.0	ights=fl: ights=fl: ights.ilo os FL_NUM	ights.fil ights.fil oc[177:18	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	:115':0}) DNTH DAN		MSP DTW		- 8	52 0. 24 0.	0 1	
	f1: f2:	ights=f1: ights=f1: ights.i1d  S  FL_NUM  2834  2839  86  87	ights.fil ights.fil ights.fil ic[177:18  MONTH  1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 9 10		MSP DTW MSP DTW	SEA JFK DTW MSP	8 17 16	.52 0. 24 0. 32 Nat 49 1.	0 1 0 0 N 1 0 0	.0 .0 .0
	f1: f1:	ights=fl: ights=fl: ights.ilco bs FL_NUM 2834 2839 86 87 423	ights.fil ights.fil ights.fil ic[177:18 MONTH 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 9 10 10		MSP DTW MSP DTW JFK	SEA JFK DTW MSP ATL	8 17 16 16	.52 0. 24 0. 32 Nat 49 1.	0 1 0 0 N 1 0 0	.0 .0 .0 .0
	f1: f2:	ights=f1: ights=f1: ights.i1d  S  FL_NUM  2834  2839  86  87	ights.fil ights.fil ights.fil ic[177:18  MONTH  1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 9 10		MSP DTW MSP DTW	SEA JFK DTW MSP	8 17 16 16	52 0. 24 0. 32 Nat 49 1. 00 0.	0 1 0 0 N 1 0 0 0 0	.0 .0 .0
	f1: f1:	ights=fl: ights=fl: ights.ilco >> FL_NUM 2834 2839 86 87 423 440	ights.fil ights.fil ic[177:18 MONTH 1 1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 9 10 10 10		MSP DTW MSP DTW JFK JFK JFK	SEA JFK DTW MSP ATL ATL	8 17 16 16 16 8	52 0. 24 0. 32 Nal 49 1. 00 0.	0 1 0 0 N 1 0 0 0 0	.0 .0 .0 .0
	f1: f1:	ights=f1: ights=f1: ights.ilc bs  FL_NUM 2834 2839 86 87 423 440 485	ights.fil ights.fil ights.fil ights.fil 177:18 MONTH 1 1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 10 10 10 10 10		MSP DTW MSP DTW JFK JFK JFK	SEA JFK DTW MSP ATL ATL SEA	8 17 16 16 16 8	.52 0. 24 0. 32 Nat 49 1. 00 0. 49 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1:	ights=f1: ights=f1: ights.ilc bs  FL_NUM 2834 2839 86 87 423 440 485	ights.fil ights.fil ights.fil ights.fil 177:18 MONTH 1 1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 10 10 10 10 10		MSP DTW MSP DTW JFK JFK JFK	SEA JFK DTW MSP ATL ATL SEA	8 17 16 16 16 8	.52 0. 24 0. 32 Nat 49 1. 00 0. 49 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=f1: ights=f1: ights.ilc bs  FL_NUM 2834 2839 86 87 423 440 485	ights.fil ights.fil phts.fil oc[177:18 MONTH 1 1 1 1 1 1	ln <b>a</b> ({'ARR_DE ln <b>a</b> ({'dep_de 5]	9 10 10 10 10 10		MSP DTW MSP DTW JFK JFK JFK	SEA JFK DTW MSP ATL ATL SEA	8 17 16 16 16 8	.52 0. 24 0. 32 Nat 49 1. 00 0. 49 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1:	ights=fl:ights=fl:ights.ilco  FL NUM 2834 2839 86 87 423 440 485 557	ights.fil ights.fil phts.fil oc[177:18 MONTH 1 1 1 1 1 1 1	lna({'ARR_DE lna(('dep_de 5]  DAY_OF_MC	DNTH DAY 9 9 10 10 10 10 10 10 10 terrows()	6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 17 16 16 16 8 19 9	.52 024 032 Nat .49 100 049 049 041 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=fl: ights=fl: ights.ilc >> FL NUM 2834 2839 86 87 423 440 485 557	ights.fil ights.fil ights.fil ic[177:18 MONTH 1 1 1 1 1 1 1 1 4 x,row ir s.loc[ir	lna({'ARR_DE lna({'dep_de 5]	DNTH DAY 9 9 10 10 10 10 10 10 10 terrows()	6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 17 16 16 16 8 19 9	.52 024 032 Nat .49 100 049 049 041 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=fl; ights=fl; ights.ile >> FL NUM 2834 2839 86 87 423 440 485 557	ights.fil ights.fil ights.fil ic[177:18 MONTH 1 1 1 1 1 1 1 1 4 x,row ir s.loc[ir	lna({'ARR_DE lna(('dep_de 5]  DAY_OF_MC	DNTH DAY 9 9 10 10 10 10 10 10 10 terrows()	6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 17 16 16 16 8 19 9	.52 024 032 Nat .49 100 049 049 041 0.	0 1 0 0 N 1 0 0 0 0	
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=flights=flights.ildights	MONTH	ina({'ARR_DE Ina(('dep_de 5] DAY_OF_MC	DNTH DAN 9 9 10 10 10 10 10 10 10 10 RR_TIME']	6 6 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 17 16 16 16 8 19 9	52 0. 24 0. 24 1. 25 Nafa 1. 26 0. 26 1. 27 0. 28 1. 29 0. 29 0. 20 0. 20 0. 20 0.	0 1 0 0 0 N 1 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=fl; ights=fl; ights.ilc 2834 2839 86 87 423 440 455 557 .mport m flights. 3.2s	MONTH	ina({'ARR_DE Ina(('dep_de 5] DAY_OF_MC	DNTH DAY 9 9 10 10 10 10 10 10 10 terrows()	6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 17 16 16 16 16 8 19 9 9 TIME']/1	52 0. 24 0. 24 1. 25 Naf 1. 26 0. 26 1. 27 0. 28 1. 29 0. 29 0. 20	DEP_DEL15	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=fl; ights=fl; ights.ilc ights.ilc 2839 86 87 423 440 485 557 mport m for inde flights. 32s FL NUM	MONTH	ina({'ARR_DE Ina(('dep_de 5] DAY_OF_MC	DNTH DAN 9 9 10 10 10 10 10 10 10 10 RR_TIME']	6 6 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK JFK MSP	SEA  JFK  DTW  MSP  ATL  ATL  SEA  DTW  CRS_ARR_  GIN DE	8 17 16 16 16 16 8 19 9 9 TIME']/1	52 0. 24 0. 24 1. 20 0. 24 1. 20 0. 249 0. 249 0. 245 1. 210 0. 210 0. 210 0. 221 0. 221 0.	DEP_DEL15	ARR_DEL15
andling Missing Data	file file file file file file file file	ights=fl; ights=fl; ights.ilc ights.ilc 2839 86 87 423 440 485 557 .mport me for linghts. 32s FL_NUM 1399 1476	MONTH	lna({'ARR_DE lna({'dep_de 5] DAY_OF_MC	DNTH DAN 9 9 10 10 10 10 10 10 10 10 10 10 10 10 10	6 6 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK MSP  OOOT (row[ 'C	SEA  JFK  DTW  MSP  ATL  ATL  SEA  DTW  RS_ARR_  ATL  SEA  DTW  ATL  SEA  DTW  ATL  SEA  DTW  ATL  SEA  DTW	8 17 16 16 16 8 19 9 9 1 TIME']/1	52 0. 24 0. 23 Nafa 49 1. 00 0. 49 0. 45 1. 12 0.  6 ARR_TIME 21	DEP_DEL15	ARR_DEL15
andling Missing Data	f1: f1: v 0: 177 178 179 180 181 182 183 184	ights=fl; ights=fl; ights.ilc ights.ilc 2839 86 87 423 440 485 557 mport m for inde flights. 32s FL NUM		lna({'ARR_DE lna({'dep_de 5] DAY_OF_MC	DNTH DAN 9 9 10 10 10 10 10 10 10 10 10 10 10 10 10	6 6 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK MSP  DOOR (row[ 'C	SEA  JFK  DTW  MSP  ATL  ATL  SEA  DTW  RS_ARR_  TRS_ARR_  TRS_ARR_  ATL  SEA  SEA  DTW  ATL  SEA  ATL  SE	8 17 16 16 16 16 8 19 9 9 TIME']/1	52 0. 24 0. 24 1. 20 0. 24 1. 20 0. 249 0. 249 0. 245 1. 210 0. 210 0. 210 0. 221 0. 221 0.	DEP_DEL15	ARR_DEL15 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.
andling Missing Data	file file file file file file file file	ights=fl; ights=fl; ights.il: 2839 86 87 423 440 485 557 mmport m flights. 32s FL_NUM 1399 1476 1597		lna({'ARR_DE lna(('dep_de 5] DAY_OF_MC	DNTH DAN 9 9 10 10 10 10 10 10 10 10 11 11 11 11 11	6 6 7 7 7 7 7 7	MSP DTW MSP DTW JFK JFK MSP  OOOT (row[ 'C	SEA JFK DTW MSP ATL SEA DTW  RS_ARR_  GIN DE ATL S DTW  ATL S SEA N  SEA N	8 17 16 16 16 16 18 19 9 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	52 0. 24 0. 24 1. 20 0. 24 1. 20 0. 24 1. 20 0. 24 1. 20 0. 21 2. 21 14 12	DEP_DEL15	ARR_DEL15 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.





	from sklearn.preprocessing import OneHotEncoder oh=OneHotEncoder() z=oh.fit_transform(flights.iloc[:,4:5]).toarray() t=oh.fit_transform(flights.iloc[:,5:6]).toarray() z  0.0s						
Data Transformation	array([[1., 0., 0., 0., 0.],						
	t						
	array([[0., 0., 0., 0., 1.],						
Feature Engineering	Attached the codes in the final submission.						
Save Processed Data	_						