



Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	740080
Project Title	Flight Delay Prediction using Machine Learning.
Maximum Marks	6 Marks

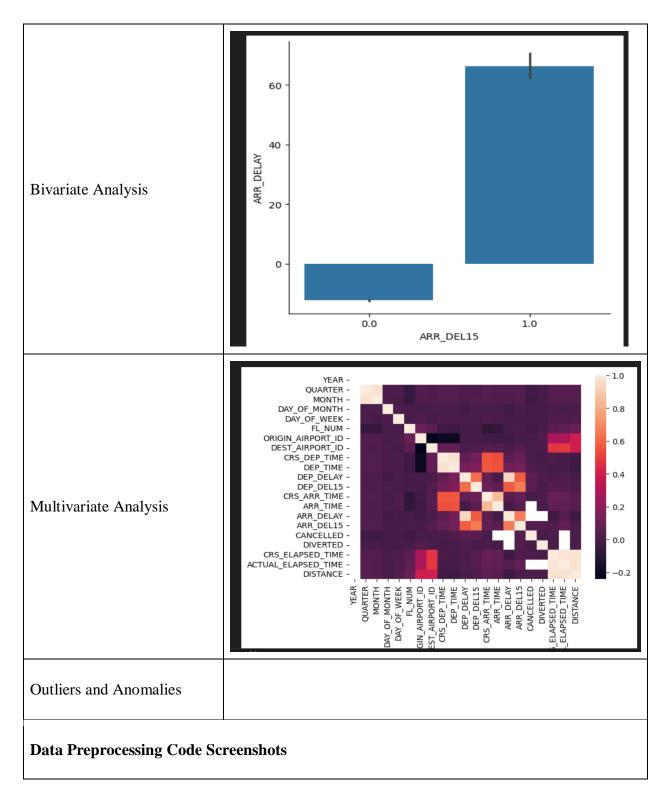
Data Exploration and Preprocessing Template

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	Descri	ption									
	YEAR	QUARTER	MONTH	DAY OF MONTH	DAY OF WEEK	FL NUM	ORIGIN AIR	PORT ID DEST	AIRPORT ID	CRS DEP TIME	DEP TIM
	count 11231.0			11231.000000		11231.000000			1231.000000		11124.00000
	mean 2016.0	2.544475	6.628973	15.790758	3.960199	1334.325617	1233	4.516695 1.	2302.274508	1320.798326	1327.18941
	std 0.0	1.090701	3.354678	8.782056	1.995257	811.875227	159	5.026510	1601.988550	490.737845	500.3064
	min 2016.0	1.000000	1.000000	1.000000	1.000000	7.000000	1039	7.000000 1	0397.000000	10.000000	1.0000
	25% 2016.0	2.000000	4.000000	8.000000	2.000000	624.000000	1039	7.000000 1	0397.000000	905.000000	905.0000
	50% 2016.0	3.000000	7.000000	16.000000	4.000000	1267.000000	1247	8.000000 1.	2478.000000	1320.000000	1324.0000
	75% 2016.0	3.000000	9.000000	23.000000	6.000000	2032.000000	1348	7.000000 1.	3487.000000	1735.000000	1739.0000
	max 2016.0	4.000000	12.000000	31.000000	7.000000	2853.000000	1474	7.000000 1-	4747.000000	2359.000000	2400.0000
Data Overview	8 rows × 21 colu										
		CRS_ARR_TIME	ARR_TIME			CANCELLED				L_ELAPSED_TIME	DISTAN
	11124.000000		11116.000000			1231.000000 1		11231.0000		11043.000000	
	0.142844	1537.312795 502.512494	1523.978499 512.536041	-2.573123 39.232521	0.124672 0.330361	0.010150 0.100241	0.006589	190.6521		179.661233	1161.031 643.683
	0.349930	2.000000	1.000000	-67.000000	0.000000	0.000000	0.000000	78.3863 93.0000		77.940399 75.000000	509.000
	. 0.000000	1130.000000	1135.000000	-19.000000	0.000000	0.000000	0.000000	127.0000		117.000000	594.000
	. 0.000000	1559.000000	1547.000000	-10.000000	0.000000	0.000000	0.000000	159,0000		149.000000	907.000
	0.000000	1952.000000	1945.000000	1,000000	0.000000	0.000000	0.000000	255.0000		236.000000	1927.0000
	. 1.000000	2359.000000	2400.000000	615.000000	1.000000	1.000000	1.000000	397.0000		428.000000	2422.0000
Univariate Analysis	f1	ights[' t.show(YEAR'].	b.pyplot			cind='p	de', aut	opct=	*%.04**)	











	flight flight ✓ 0.1s	ts=pd.read_ ts	Cov(IIIgii									
	Y	EAR QUAF	TER MON	TH DAY_OF_MO	NTH DAY_O	F_WEEK UNIQ	UE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN_AIRPORT_I	D ORIGIN	. DEP_DEL15
		2016						N836DN	1399	1039		. 0.0
		2016 2016					DL DL	N964DN N813DN	1476 1597	1143		. 0.0
		2016					DL	N587NW	1768	1474		. 0.0
	4 2	2016						N836DN	1823	1474	47 SEA	. 0.0
			4		30		 DL	 N940DL	1715	114	 33 DTW	0.0
		2016		12	30		DL	N836DN		1474		. 1.0
		2016						N583NW	1823	1143		. 0.0
	11229 2	2016		12				N554NW	1901	1039	97 ATL	. 0.0
oading Data												
ading Data												
	S_ARR_	TIME AR 2143	R_TIME . 2102.0	ARR_DELAY 4	ARR_DEL15 0.0	CANCELLED		D CRS_E	LAPSED_TI	ME ACTUAL_E 338	LAPSED_TIME 295.0	
		1435	1439.0	-41.0 4.0	0.0			0		110	115.0	
		1215	1142.0	-33.0	0.0					335	300.0	
			1345.0									
		607	615.0	8.0						247	259.0	1927
		1223	1148.0	-35.0	0.0			0		138	105.0	
		2046 2210	2100.0 2154.0	14.0 -16.0	0.0			0		201 311	181.0 295.0	
		1806	1801.0							336	332.0	
		925	913.0	-12.0	0.0	(0			120	110.0	594
		ghts=fli	ghts.fill	.n a ({'ARR_DE	L15':1})						110.0	334
	fli	ghts=fli ghts=fli ghts.ilo	ghts.fill	.na({'ARR_DE .na({'dep_de	L15':1})					-	110.0	34
	fli fli ✓ 0.0	ghts=fli ghts=fli ghts.ilo s FL_NUM	ghts.fill	.na({'ARR_DE .na({'dep_de	L15':1}) l15':0})	_OF_WEEK	ORIGIN I		S_ARR_TIM	ME DEP_DEL15	6 ARR_DEL1	5
	fli fli <- 0.0	ghts=fli ghts=fli ghts.ilo	ghts.fill ghts.fill c[177:185	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) l15':0})		ORIGIN I MSP	DEST CR: SEA JFK	S_ARR_TIM 8	//E DEP_DEL15 52 0.0	6 ARR_DEL1	.5
	fli fli ✓ 0.0	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834	ghts.fill ghts.fill c[177:185	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) DNTH DAY	_OF_WEEK 6	ORIGIN I MSP DTW	SEA	S_ARR_TIM	ME DEP_DEL15 52 0.0 24 0.0	6 ARR_DEL1 0 1 0 0	5
	fli fli	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87	ghts.fill ghts.fill c[177:185 MONTH 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) DNTH DAY 9 9 10	COF_WEEK 6 6 7 7	ORIGIN I MSP DTW MSP DTW	SEA JFK DTW MSP	5_ARR_TIM 8 17. 16 16	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0	S ARR_DEL1) 1.) 0 I 1.	5 0 0 0
	fli fli 0.0 177 178 179 180 181	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423	ghts.fill ghts.fill c[177:185 MONTH 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) NTH DAY 9 10 10	_ OF_WEEK 6 6 7 7 7 7	ORIGIN I MSP DTW MSP DTW JFK	SEA JFK DTW MSP ATL	5_ARR_TIM 8 17. 16 16 16	ME DEP_DEL15 52 0.0 32 NaN 49 1.0 00 0.0	6 ARR_DEL1 10 0 11 1 10 0 10 0	5 0 0 0
	fli fli v 0.0 177 178 179 180 181	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423 440	ghts.fill [177:188 MONTH 1 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':e}) NNTH DAY 9 10 10 10	COF_WEEK 6 6 7 7	ORIGIN I MSP DTW MSP DTW JFK JFK	SEA JFK DTW MSP ATL ATL	5_ARR_TIM 8 17. 16 16 16 8	ME DEP_DEL15 52 0.0 24 0.0 32 Nan 49 1.0 00 0.0	5 ARR_DEL1 1 0 0 1 1 0 0 0 0 0 0 0 0 0	5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	fli fli 0.0 177 178 179 180 181	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423	ghts.fill ghts.fill c[177:185 MONTH 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) NTH DAY 9 10 10	_ OF_WEEK 6 6 7 7 7 7 7	ORIGIN I MSP DTW MSP DTW JFK JFK	SEA JFK DTW MSP ATL	5_ARR_TIM 8 17: 16 16 16 8 19:	ME DEP_DEL15 52 0.0 24 0.0 32 Nan 49 1.0 00 0.0	5 ARR DEL1	5 0 0 0
	fli fli 0.0- 177 178 179 180 181 182 183	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423 440 485	ghts.fill ghts.fill c[177:185 MONTH 1 1 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) NNTH DAY 9 10 10 10 10 10	OF WEEK 6 6 7 7 7 7 7 7	ORIGIN I MSP DTW MSP DTW JFK JFK	SEA JFK DTW MSP ATL ATL SEA	5_ARR_TIM 8 17: 16 16 16 8 19:	AE DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 00 49 49 1.0	5 ARR DEL1	5 .0 .0 .0 .0
andling Missing Data	fli fli 0.0- 177 178 179 180 181 182 183	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423 440 485	ghts.fill ghts.fill c[177:185 MONTH 1 1 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) NNTH DAY 9 10 10 10 10 10	OF WEEK 6 6 7 7 7 7 7 7	ORIGIN I MSP DTW MSP DTW JFK JFK	SEA JFK DTW MSP ATL ATL SEA	5_ARR_TIM 8 17: 16 16 16 8 19:	AE DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 00 49 49 1.0	5 ARR DEL1	5 .0 .0 .0 .0
andling Missing Data	fii fii v 00 177 178 179 180 181 182 183 184	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423 440 485 557	month 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	L15':1}) 115':0}) NNTH DAY 9 10 10 10 10 10	OF WEEK 6 6 7 7 7 7 7 7	ORIGIN I MSP DTW MSP DTW JFK JFK	SEA JFK DTW MSP ATL ATL SEA	5_ARR_TIM 8 17: 16 16 16 8 19:	AE DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 00 49 49 1.0	5 ARR DEL1	5 .0 .0 .0 .0
andling Missing Data	fii fili v 00 177 178 179 180 181 182 183 184	ghts=fli ghts=fli ghts.ilo s FL_NUM 2834 2839 86 87 423 440 485 557	ghts.fill ghts.fill c[177:188 MONTH 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	.na({ 'ARR_DE .na({ 'dep_de :]	DNTH DAY 9 10 10 10 10 10 10 terrows()	OF WEEK 6 6 7 7 7 7 7 7 7 7 7 7 7 :	ORIGIN I MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	5 ARR TIM 8 17: 16 16: 16: 8 19: 9	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 49 0.0 45 1.0	5 ARR DEL1 0 1 0 0 1 1 0 0 0 0 0 0 0 0	5 .0 .0 .0 .0
andling Missing Data	fii fili v 00 177 178 179 180 181 182 183 184	ghts=flightside flightside flight	month mo	na({'ARR_DE na(('dep_de)) DAY_OF_MO	DNTH DAY 9 10 10 10 10 10 10 terrows()	OF WEEK 6 6 7 7 7 7 7 7 7 7 7 7 7 :	ORIGIN I MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	5 ARR TIM 8 17: 16 16: 16: 8 19: 9	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 49 0.0 45 1.0	5 ARR DEL1 0 1 0 0 1 1 0 0 0 0 0 0 0 0	5 .0 .0 .0 .0
andling Missing Data	fii fii 7 00 177 178 179 180 181 182 183 184	ghts=flightside flightside flight	month mo	na({'ARR_DE na(('dep_de]) DAY_OF_MO flights.it dex,'CRS_Al	L15':1}) 115':0}) NNTH DAY 9 9 10 10 10 10 10 10 RR_TIME']	OF WEEK 6 6 7 7 7 7 7 7 7 7 7 7 7 :	ORIGIN I MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW	8 ARR TIM 8 17, 16, 16, 16, 18, 19, 9	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 49 0.0 45 1.0 12 0.0	5 ARR DEL1 0 1 0 0 1 1 0 0 0 0 0 0 0 0	5 0 0 0 0 0 0 0
andling Missing Data	fii fii 7 00 177 178 179 180 181 182 183 184	ghts=fli ghts=fli ghts.ilo s FNNM 2834 2839 86 87 423 440 485 557 mmport mm or index flights.i	MONTH:	na({'ARR_DE na(('dep_de]) DAY_OF_MO flights.it dex,'CRS_Al	L15':1}) 115':0}) NNTH DAY 9 9 10 10 10 10 10 10 RR_TIME']	_OF_WEEK	ORIGIN I MSP DTW MSP DTW JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW RS_ARR_ GIN DE	8 ARR TIM 8 17, 16, 16, 16, 18, 19, 9	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 49 0.0 45 1.0 12 0.0	i ARR_DEL1 1	5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
andling Missing Data	fii fiii v 00 177 178 180 181 182 183 184	ghts=fli ghts=fli ghts.ilo 5 5 5 1 2839 86 87 423 440 485 557 557 mport midel flights.il	ghts.fill filtrian MONTH 1 1 1 1 1 1 1 1 Month filth fil	na({'ARR_DE na(('dep_de]) DAY_OF_MO flights.it dex,'CRS_Al	L15':1}) 115':0}) NTH DAY 9 9 10 10 10 10 10 10 RR_TIME']	_OF_WEEK	ORIGIN I MSP DTW MSP DTW JFK JFK JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW RS_ARR_ GIN DE	8 ARR TIM 8 17, 16 16 16 16 8 19 9 9	ME DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 49 0.0 45 1.0 12 0.0	S ARR_DELT 1	ARR_DEL15
andling Missing Data	filification of the state of th	ghts=fli ghts=fli ghts:ilo* 2834 2839 86 87 423 440 485 557 mmport mm flights.l 1399 1476	ghts.fill this.fill this.fill this.fill this this this this this this this this	ns({'ARR_DE ns(('dep_de]) DAY_OF_MO flights.ii dex,'CRS_AI	L15':1)) 115':0)) ONTH DAY 9 9 10 10 10 10 10 10 10 MONTH 1	_OF_WEEK	ORIGIN I MSP DTW MSP DTW JFK JFK JFK JFK MSP	SEA JFK DTW MSP ATL ATL SEA DTW RS_ARR_ GIN DE ATL S DTW MSP	8 ARR TIM 8 177 166 160 160 188 199 9 9	#E DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 60 49 0.0 45 1.0 60 60 60 60 60 60 60 61 62 61 62 61 63 64 64 65 66 66 66 67 68 68 68 68 68 68 68 68 68 68 68 68 68	5 ARR DEL1 1 1 0 0 1 1 0 0 0 0 1 1 DEP_DEL15 0.0	ARR_DEL15
andling Missing Data	filification of the state of th	ghts=fli ghts=fli ghts=fli ghts.ilo' 2834 2839 86 87 423 440 485 557 mmport mi 11ghts.l 13ges 1476 1597	ghts.fill ths.fill th	ns({'ARR_DE ns(('dep_de]) DAY_OF_MO flights.id dex,'CRS_AI	L15':1)) 115':0)) INTH DAY 9 9 10 10 10 10 10 10 10 10 10 10 10 10 10	_OF_WEEK	ORIGIN I MSP DTW MSP DTW JFK JFK MSP OO (row['C	SEA JFK DTW MSP ATL ATL SEA DTW RS_ARR_ T GIN DE ATL S DTW ATL S AT	8 ARR TIM 8 177 166 166 8 199 9 TTIME']/1	#E DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 00 0.0 45 1.0 12 0.0 ### Out to the control of the con	5 ARR_DEL15 0 1 0 0 0 0 0 0 0 0 0 1	ARR_DEL15 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.
andling Missing Data	filification of the state of th	ghts=fli ghts=fli ghts:ilo* 2834 2839 86 87 423 440 485 557 mmport mm flights.l 1399 1476	ghts.fill ths.fill th	ns({'ARR_DE ns(('dep_de]) DAY_OF_MO flights.it dex, 'CRS_Al H DAY_OF_1	L15':1)) 115':0)) INTH DAY 9 10 10 10 10 10 10 10 10 10 10 11 11 11	_OF_WEEK	ORIGIN I MSP DTW MSP DTW JFK JFK MSP OOT (row['C' C'	SEA JFK DTW MSP ATL ATL SEA DTW RS_ARR_ T GIN DE ATL S DTW ATL S SEA ATL S	8 ARR TIM 8 177 166 160 160 188 199 9 9	#E DEP_DEL15 52 0.0 24 0.0 32 NaN 49 1.0 60 49 0.0 45 1.0 60 60 60 60 60 60 60 61 62 61 62 61 63 64 64 65 66 66 66 67 68 68 68 68 68 68 68 68 68 68 68 68 68	5 ARR_DEL15 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ARR_DEL15 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.





Data Transformation	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 7						
	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	_						