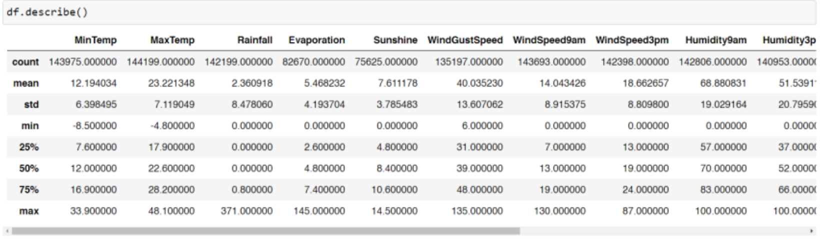
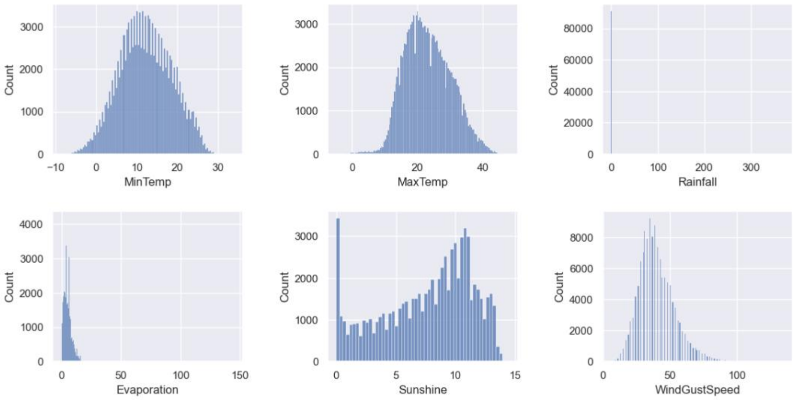
Data Exploration and Preprocessing Report

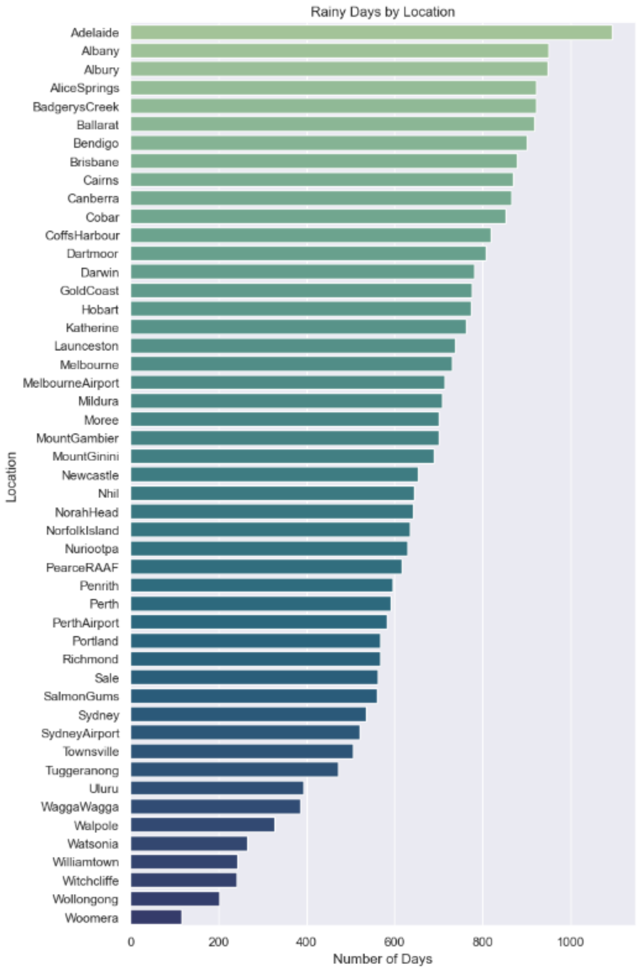
Date	15 April 2024
Team ID	Team-738164
Project Title	Rainfall Prediction Using Machine Learning
Maximum Marks	6 Marks

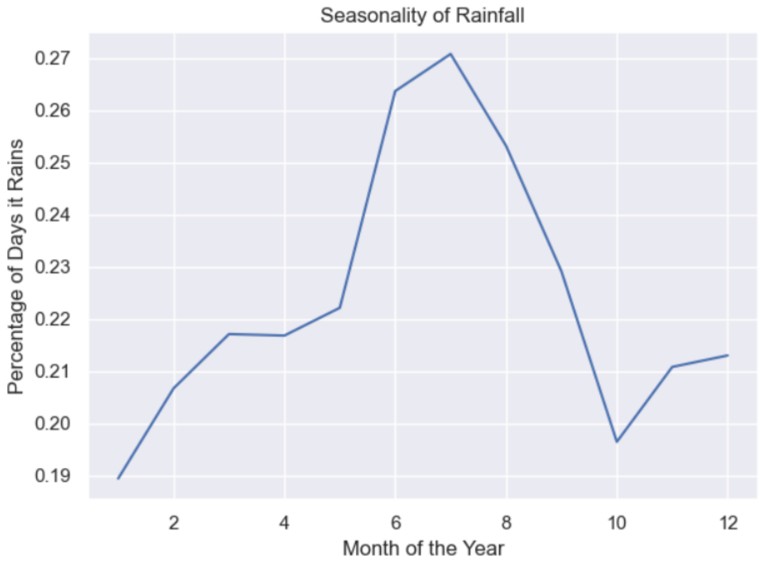
Data Exploration Screenshots:

Section	Description
Data Overview	Dimensions: 145460 rows x 23 columns Descriptive Statistics: 
Univariate Analysis	
Bivariate Analysis	

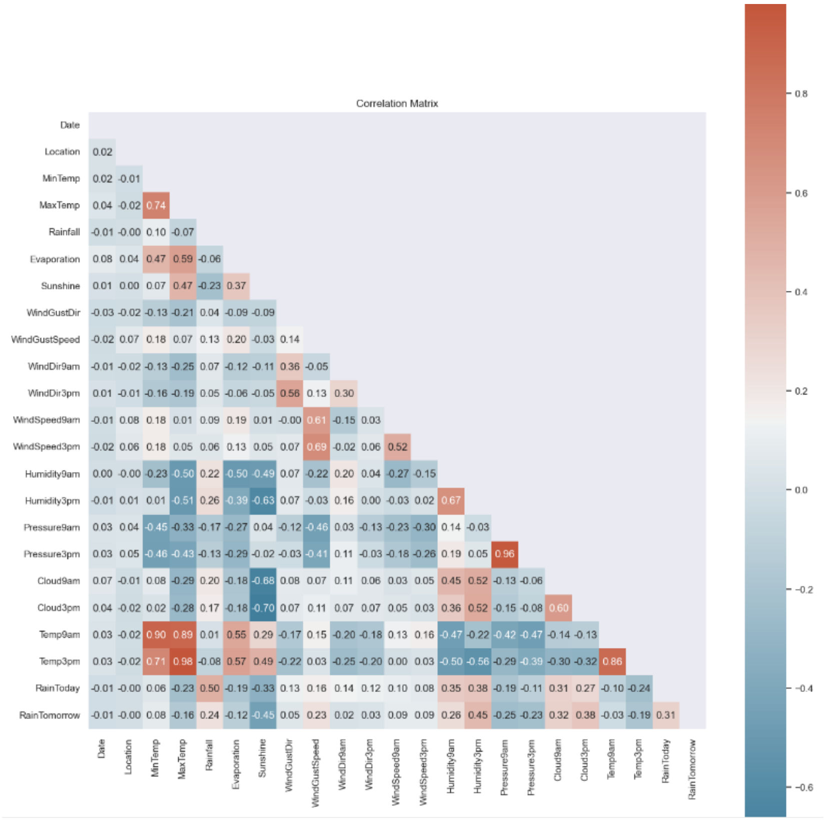
```
df.describe()
```

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3p
count	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000	135197.000000	143693.000000	142398.000000	142806.000000	140953.0000
mean	12.194034	23.221348	2.360918	5.468232	7.611178	40.035230	14.043426	18.662657	68.880831	51.5391
std	6.398495	7.119049	8.478060	4.193704	3.785483	13.607062	8.915375	8.809800	19.029164	20.7959
min	-8.500000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	0.000000	0.0000
25%	7.600000	17.900000	0.000000	2.600000	4.800000	31.000000	7.000000	13.000000	57.000000	37.0000
50%	12.000000	22.600000	0.000000	4.800000	8.400000	39.000000	13.000000	19.000000	70.000000	52.0000
75%	16.900000	28.200000	0.800000	7.400000	10.600000	48.000000	19.000000	24.000000	83.000000	66.0000
max	33.900000	48.100000	371.000000	145.000000	14.500000	135.000000	130.000000	87.000000	100.000000	100.0000


Rainy Days by Location
Location
Adelaide
Albany
Albury
AliceSprings
BadgerysCreek
Ballarat
Bendigo
Brisbane
Cairns
Canberra
Cobar
CoffsHarbour
Dartmoor
Darwin
GoldCoast
Hobart
Katherine
Launceston
Melbourne
MelbourneAirport
Mildura
Moree
MountGambier
MountGinini
Newcastle
Nhil
NorahHead
NorfolkIsland
Nuriootpa
PearceRAAF
Penrith
Perth
PerthAirport
Portland
Richmond
Sale
SalmonGums
Sydney
SydneyAirport
Townsville
Tuggeranong
Uluru
WaggaWagga
Walpole
Watsonia
Williamtown
Witchcliffe
Wollongong
Woomera
0
200
400
600
800
1000
Number of Days


Seasonality of Rainfall
Percentage of Days it Rains
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20
0.19
2
4
6
8
10
12
Month of the Year

Multivariate Analysis


Correlation Matrix

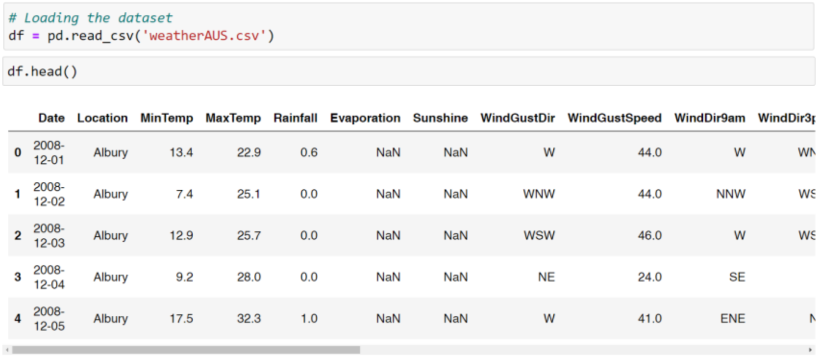
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
Date																							
Location	0.02																						
MinTemp	0.02	-0.01																					
MaxTemp	0.04	-0.02	0.74																				
Rainfall	-0.01	-0.00	0.10	-0.07																			
Evaporation	0.08	0.04	0.47	0.59	-0.06																		
Sunshine	0.01	0.00	0.07	0.47	-0.23	0.37																	
WindGustDir	-0.03	-0.02	-0.13	-0.21	0.04	-0.09	-0.09																
WindGustSpeed	-0.02	0.07	0.18	0.07	0.13	0.20	-0.03	0.14															
WindDir9am	-0.01	-0.02	-0.13	-0.25	0.07	-0.12	-0.11	0.36	-0.05														
WindDir3pm	0.01	-0.01	-0.16	-0.19	0.05	-0.06	-0.05	0.56	0.13	0.30													
WindSpeed9am	-0.01	0.08	0.18	0.01	0.09	0.19	0.01	-0.00	0.61	-0.15	0.03												
WindSpeed3pm	-0.02	0.06	0.18	0.05	0.06	0.13	0.05	0.07	0.69	-0.02	0.06	0.52											
Humidity9am	0.00	-0.00	-0.23	-0.50	0.22	-0.50	-0.49	0.07	-0.22	0.20	0.04	-0.27	-0.15										
Humidity3pm	-0.01	0.01	0.01	-0.51	0.26	-0.39	-0.63	0.07	-0.03	0.16	0.00	-0.03	0.02	0.67									
Pressure9am	0.03	0.04	-0.45	-0.33	-0.17	-0.27	0.04	-0.12	-0.46	0.03	-0.13	-0.23	-0.30	0.14	-0.03								
Pressure3pm	0.03	0.05	-0.46	-0.43	-0.13	-0.29	-0.02	-0.03	-0.41	0.11	-0.03	-0.18	-0.26	0.19	0.05	0.96							
Cloud9am	0.07	-0.01	0.08	-0.29	0.20	-0.18	-0.68	0.08	0.07	0.11	0.06	0.03	0.05	0.45	0.52	-0.13	-0.06						
Cloud3pm	0.04	-0.02	0.02	-0.28	0.17	-0.18	-0.70	0.07	0.11	0.07	0.07	0.05	0.03	0.36	0.52	-0.15	-0.08	0.60					
Temp9am	0.03	-0.02	0.90	0.89	0.01	0.55	0.29	-0.17	0.15	-0.20	-0.18	0.13	0.16	-0.47	-0.22	-0.42	-0.47	-0.14	-0.13				
Temp3pm	0.03	-0.02	0.71	0.98	-0.08	0.57	0.49	-0.22	0.03	-0.25	-0.20	0.00	0.03	-0.50	-0.56	-0.29	-0.39	-0.30	-0.32	0.86			
RainToday	-0.01	-0.00	0.06	-0.23	0.50	-0.19	-0.33	0.13	0.16	0.14	0.12	0.10	0.08	0.35	0.38	-0.19	-0.11	0.31	0.27	-0.10	-0.24		
RainTomorrow	-0.01	-0.00	0.08	-0.16	0.24	-0.12	-0.45	0.05	0.23	0.02	0.03	0.09	0.09	0.26	0.45	-0.25	-0.23	0.32	0.38	-0.03	-0.19	0.31	

Outliers and Anomalies

1. Multiple columns have clear outliers (e.g., the max Rainfall value is 371.0 despite the 75th percentile being 0.8)
2. Not seeing any values that are immediate cause for concern (such as a negative value for minimum Rainfall)

Data Preprocessing Code Screenshots:

Loading Data

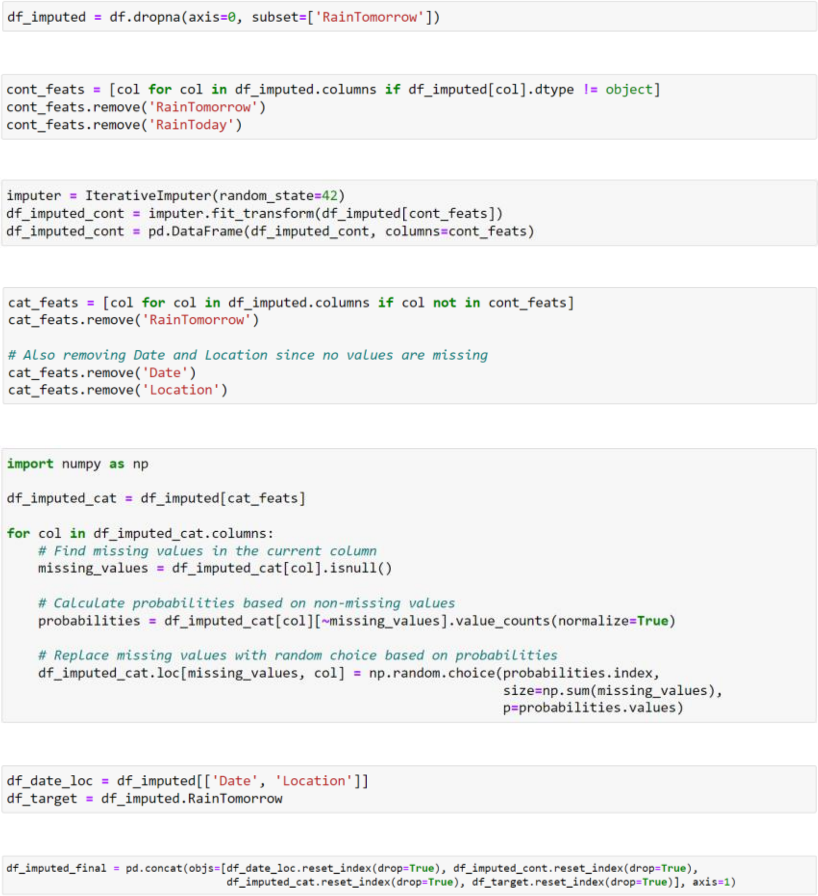


```
# Loading the dataset
df = pd.read_csv('weatherAUS.csv')
```

```
df.head()
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3p
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WN
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WS
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WS
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	N

Handling Missing Data



```
df_imputed = df.dropna(axis=0, subset=['RainTomorrow'])
```

```
cont_feats = [col for col in df_imputed.columns if df_imputed[col].dtype != object]
cont_feats.remove('RainTomorrow')
cont_feats.remove('RainToday')
```

```
imputer = IterativeImputer(random_state=42)
df_imputed_cont = imputer.fit_transform(df_imputed[cont_feats])
df_imputed_cont = pd.DataFrame(df_imputed_cont, columns=cont_feats)
```

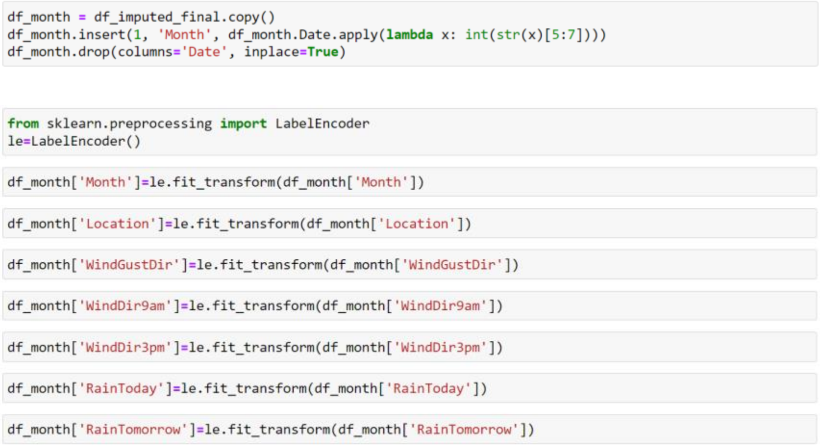
```
cat_feats = [col for col in df_imputed.columns if col not in cont_feats]
cat_feats.remove('RainTomorrow')

# Also removing Date and Location since no values are missing
cat_feats.remove('Date')
cat_feats.remove('Location')
```

```
import numpy as np

df_imputed_cat = df_imputed[cat_feats]

for col in df_imputed_cat.columns:
    # Find missing values in the current column
    missing_values = df_imputed_cat[col].isnull()

    # Calculate probabilities based on non-missing values
    probabilities = df_imputed_cat[col][~missing_values].value_counts(normalize=True)

    # Replace missing values with random choice based on probabilities
    df_imputed_cat.loc[missing_values, col] = np.random.choice(probabilities.index,
                                                               size=np.sum(missing_values),
                                                               p=probabilities.values)
```

```
df_date_loc = df_imputed[['Date', 'Location']]
df_target = df_imputed.RainTomorrow
```

```
df_imputed_final = pd.concat(objs=[df_date_loc.reset_index(drop=True), df_imputed_cont.reset_index(drop=True),
                                   df_imputed_cat.reset_index(drop=True), df_target.reset_index(drop=True)], axis=1)
```

Data Transformation



```
df_month = df_imputed_final.copy()
df_month.insert(1, 'Month', df_month.Date.apply(lambda x: int(str(x)[5:7])))
df_month.drop(columns='Date', inplace=True)
```

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

```
df_month['Month']=le.fit_transform(df_month['Month'])
```

```
df_month['Location']=le.fit_transform(df_month['Location'])
```

```
df_month['WindGustDir']=le.fit_transform(df_month['WindGustDir'])
```

```
df_month['WindDir9am']=le.fit_transform(df_month['WindDir9am'])
```

```
df_month['WindDir3pm']=le.fit_transform(df_month['WindDir3pm'])
```

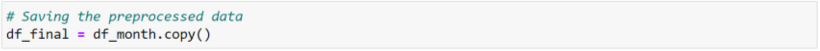
```
df_month['RainToday']=le.fit_transform(df_month['RainToday'])
```

```
df_month['RainTomorrow']=le.fit_transform(df_month['RainTomorrow'])
```

Feature Engineering

Attached the codes in final submission.

Save Processed Data



```
# Saving the preprocessed data
df_final = df_month.copy()
```