Samsung Innovation Campus

Gözde Gözütok
Hale Afra İriş
Özge Özkaya
Toyan Ünal

Session Details

- Dataset
- Data cleaning
- Data preprocessing
- Models
- Preliminary Outputs

Data collection

- Türkiye Hava Durumu ve İklim Verisi (National Centers for Environmental Information)
  - `https://geographic.org/global_weather`
- Türkiye Yangın Verisi (NASA)

Columns

City	City name from Turkey
Year	Year of the record
Month	Month of the record
Day	Day of the record
Temperature avarage, Fahrenheit / Celsius	The average temperature of the daily record in Fahrenheit and Celcius
Temperature maximum, Fahrenheit / Celsius	The maximum temperature of the daily record in Fahrenheit and Celcius
Temperature minimum, Fahrenheit / Celsius	The minimum temperature of the daily record in Fahrenheit and Celcius
Precipitation in / cm	The precipitation of the daily record in cantimeter
Average daily wind speed (miles/hour)	The average wind of the daily record in miles and hours
Visibility in miles	The visibility in miles of the daily record
Dew point, Fahrenheit / Celsius	The dew point of the daily record in Fahrenheit and Celcius
Maximum sustained wind speed (miles/hour)	The maximum sustained wind speed of the daily record in miles and hour
Maximum wind gust reported for the day	The maximum wind gust of the daily record
Sea level pressure in mBARs	The sea level of the daily record in mBARs
Station pressure in mBARs	Station pressure of the daily record in mBARs
Snow depth in inches	The snow depth of the daily record in inches
Weather type	The wheather type

Descriptive statistics of data

```
In [5]: turkey_daily_climate.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 83912 entries, 0 to 83911
Data columns (total 18 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Unnamed: 0                              83912 non-null  int64
 1   City                                    83912 non-null  object
 2   Year                                    83912 non-null  int64
 3   Month                                   83912 non-null  object
 4   Day                                     83912 non-null  int64
 5   Temperature avarage, Fahrenheit / Celsius  83912 non-null  object
 6   Temperature maximum, Fahrenheit / Celsius  83912 non-null  object
 7   Temperature minimum, Fahrenheit / Celsius  83912 non-null  object
 8   Precipitation in / cm                   37278 non-null  object
 9   Average daily wind speed (miles/hour)   80462 non-null  object
 10  Visibility in miles                     80397 non-null  object
 11  Dew point, Fahrenheit / Celsius         83912 non-null  object
 12  Maximum sustained wind speed (miles/hour)  80462 non-null  object
 13  Maximum wind gust reported for the day  3149 non-null   object
 14  Sea level pressure in mBARs             1 non-null      object
 15  Station pressure in mBARs               1 non-null      object
 16  Snow depth in inches                    1 non-null      object
 17  Weather type                            17281 non-null  object
dtypes: int64(3), object(15)
memory usage: 11.5+ MB
```

```
In [10]: turkey_daily_climate.head(2)
```

Out[10]:

	City	Year	Month	Day	Temperature avarage, Fahrenheit / Celsius	Temperature maximum, Fahrenheit / Celsius	Temperature minimum, Fahrenheit / Celsius	Precipitation in / cm	Average daily wind speed (miles/hour)	Visibility in miles	Dew point, Fahrenheit / Celsius	Maximum sustained wind speed (miles/hour)	Maximum wind gust reported for the day	Weather type
0	Adana	2010	01,02	1	57.0/13.9	66.2/19.0	44.6/7.0	NaN	3.8	6.0	47.8/8.8	10.24	NaN	NaN
1	Adana	2010	01,02	2	59.2/15.1	66.2/19.0	51.8/11.0	NaN	4.6	6.0	53.2/11.8	16.11	NaN	Rain/Drizzle

- City sütunundaki 'Agri', 'Balikesir' benzeri yazımı yanlış olan şehirler düzeltildi.
- Year Month ve Day sütunları YYYY-MM-DD formatında birleştirelerek yeni bir Date sütunu oluşturuldu.
- Temperature avarage Fahrenheit ve Celcius olarak iki değer ve string formantındaydı.
  Split ile sadece Celsius ve float formatında yeni sütun oluşturuldu.
- Aynı süreç Temperature maximum ve Temperature minimum için de uygulandı.
- Precipitation in/cm sütunu da aynı şekilde inç ve santimetre olarak iki değer ('0.01/0.03') içeriyordu.
  Cm kullanılarak yeni sütun oluşturuldu ve float tipine dönüştürüldü.
- Dew
  point Fahrenheit ve Celcius olarak iki değer ve string formantındaydı. Aynı süreç uygulanarak sadece Celsius içeren float bir sütun oluşturuldu.

Missing Value

- Yağış sütunundaki eksik veriler için Weather Type ve Şehir sütunlarına göre gruplama yapılarak. Eksik olan yerler mean değerleri ile dolduruldu.

```
In [61]: df.groupby(['Weather type','City']).PrecipitationCm.describe()
```

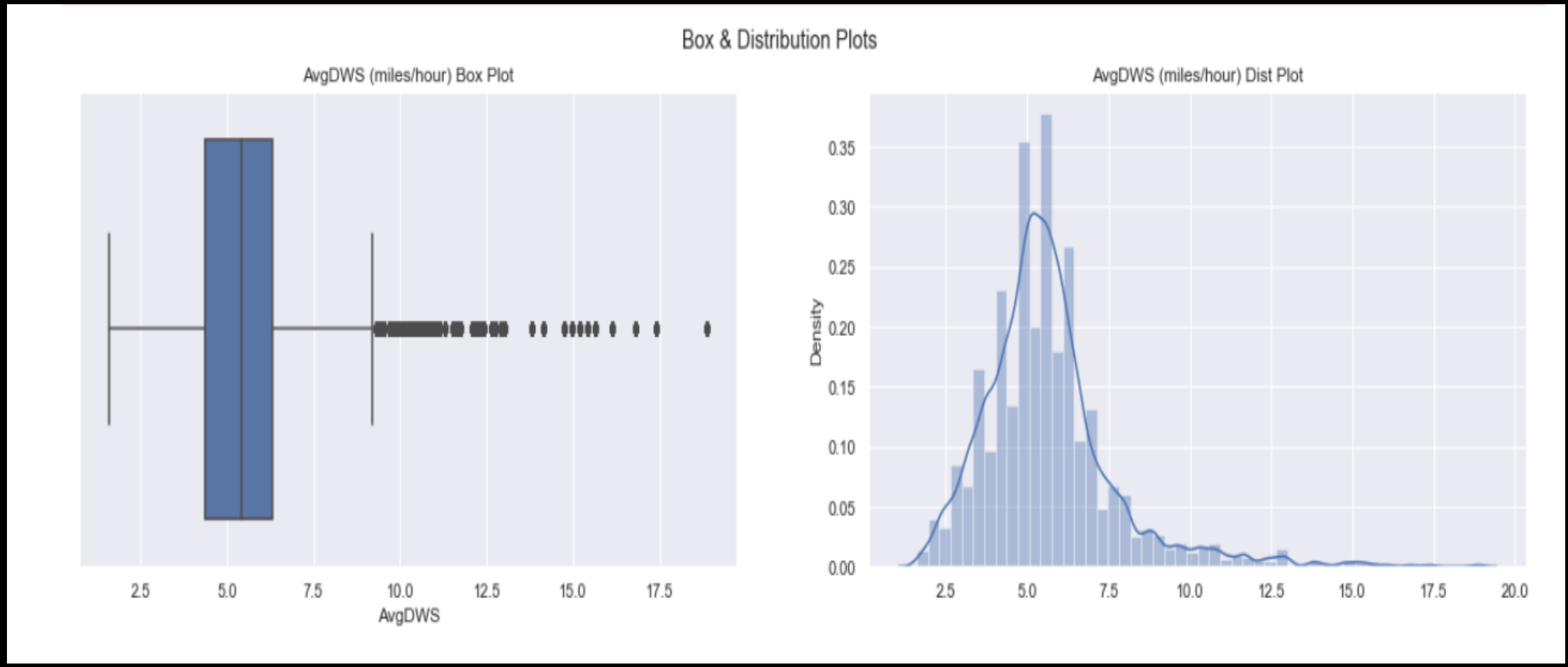
Out[61]:

		count	mean	std	min	25%	50%	75%	max
Weather type	**City**								
Fog	**Adana**	10.0	0.023000	0.072732	0.00	0.00	0.00	0.00	0.23
	Adiyaman	10.0	0.023000	0.072732	0.00	0.00	0.00	0.00	0.23
	Afyonkarahisar	13.0	0.017692	0.063791	0.00	0.00	0.00	0.00	0.23
	Agri	10.0	0.023000	0.072732	0.00	0.00	0.00	0.00	0.23
	Amasya	10.0	0.023000	0.072732	0.00	0.00	0.00	0.00	0.23
...	**...**	...	...	...	...	...	...	...	...
Thunder	**Tunceli**	1.0	0.410000	NaN	0.41	0.41	0.41	0.41	0.41
	Usak	1.0	0.410000	NaN	0.41	0.41	0.41	0.41	0.41
	Van	1.0	0.410000	NaN	0.41	0.41	0.41	0.41	0.41
	Yozgat	1.0	0.410000	NaN	0.41	0.41	0.41	0.41	0.41
	Zonguldak	1.0	0.410000	NaN	0.41	0.41	0.41	0.41	0.41

552 rows × 8 columns

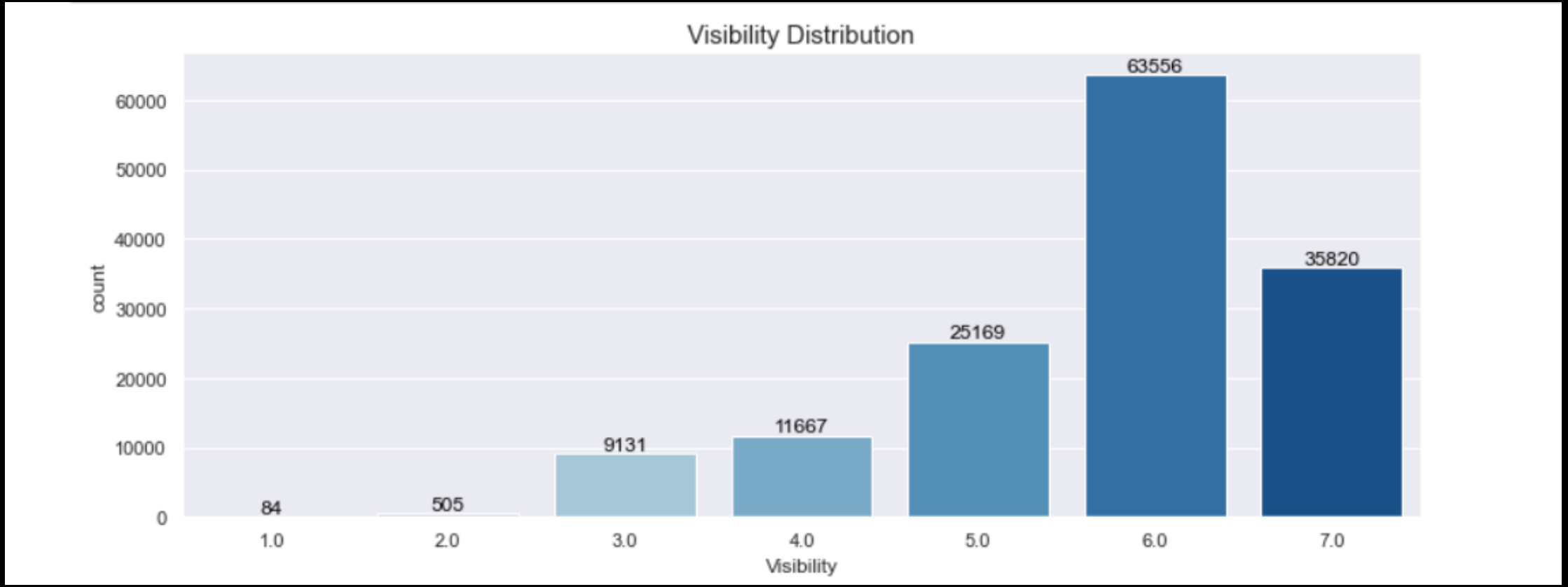
Missing Value

- Average Daily Wind Speed Şehirlere göre gruplanıp, ortalama değerleri ile dolduruldu.



Missing Value

- Visibility hem genel hem de şehirler bazında incelendiğinde hepsinde en çok tekrar eden değer 6.0 olduğu için, bu yaklaşımla eksik değerler doldurulduu.



Maximum sustained wind speed (miles/hour)
Maximum wind gust reported for the day

- Maksimum sürekli rüzgar hızı ve maksimum rüzgar esintisi, yine her şehirdeki ortalama maksimum sürekli rüzgar hızı değerleri ve ortalama maksimum rüzgar esintisi kullanılarak dolduruldu.

```
In [117]: df.groupby(['City']).median()
```

Out[117]:

City	Year	AvgDWS	Visibility	MaxSusWS	Maximum wind gust reported for the day	TempAvgCel	TempMaxCel	TempMinCel	PrecipitationCm	DewPntCel
Adana	2012.5	5.41	6.0	13.81	29.92	22.3	28.60	17.0	0.0	14.4
Adiyaman	2012.5	5.41	6.0	12.77	29.92	21.0	28.00	15.7	0.0	12.9
Afyonkarahisar	2012.5	5.41	6.0	13.81	28.88	23.2	29.00	17.8	0.0	15.1
Agri	2012.5	5.41	6.0	13.81	29.92	21.3	28.00	16.0	0.0	13.0
Amasya	2012.5	5.41	6.0	12.77	29.92	22.1	28.30	16.0	0.0	13.7
...	...	...	...	...	...	...	...	...	...	...
Tunceli	2012.5	5.41	6.0	13.81	28.88	23.1	29.00	17.3	0.0	14.8
Usak	2012.5	5.41	6.0	13.81	29.40	21.9	28.00	16.0	0.0	13.7
Van	2012.5	5.41	6.0	12.77	28.88	21.4	28.00	16.0	0.0	13.1
Yozgat	2012.5	5.41	6.0	13.81	28.88	23.3	29.05	17.8	0.0	15.1
Zonguldak	2012.5	5.41	6.0	12.77	29.92	20.9	27.40	15.0	0.0	12.9

69 rows × 10 columns

```
In [118]: df.groupby(['City']).MaxSusWS.median().value_counts()
```

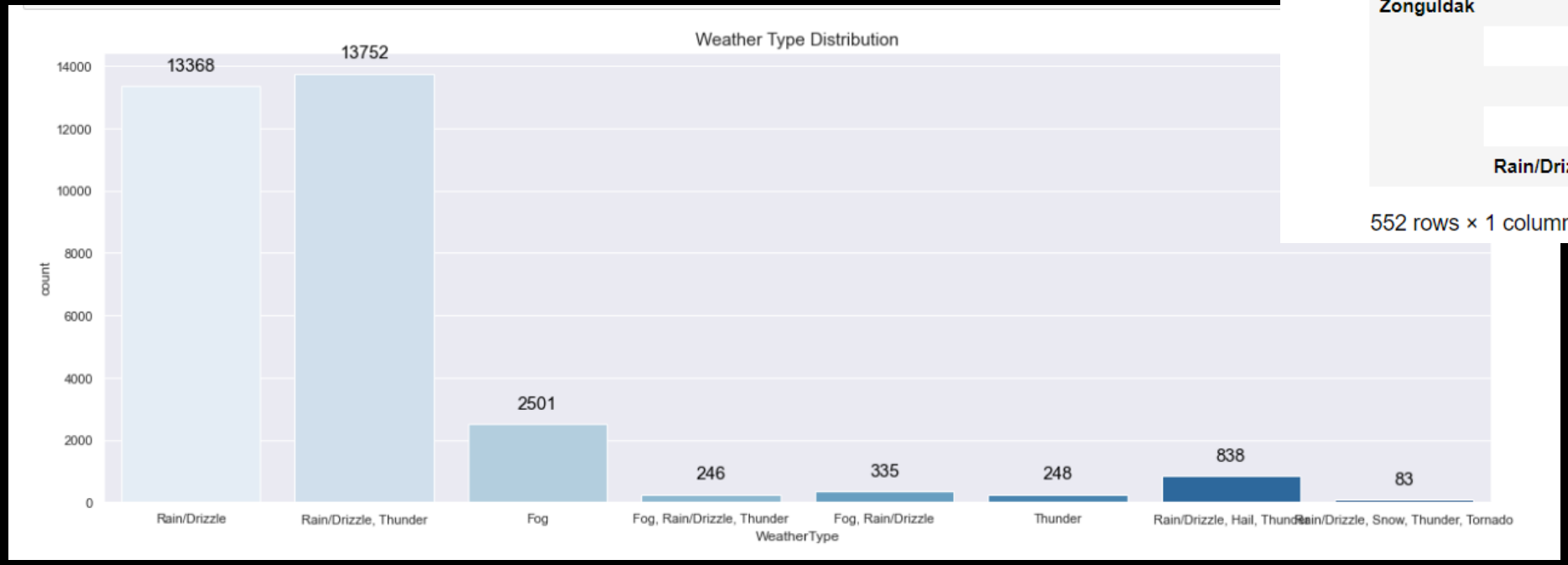
```
Out[118]: 12.77    38
          13.81    31
          Name: MaxSusWS, dtype: int64
```

Weather Type

- Her şehirde en çok tekrar eden tip ile dolduruldu

City	WeatherType	WeatherType
Adana	Rain/Drizzle, Thunder	196
	Rain/Drizzle	191
	Fog	31
	Rain/Drizzle, Hail, Thunder	11
	Fog, Rain/Drizzle	4
...	...	...
Zonguldak	Rain/Drizzle, Hail, Thunder	10
	Fog, Rain/Drizzle	5
	Thunder	5
	Fog, Rain/Drizzle, Thunder	4
	Rain/Drizzle, Snow, Thunder, Tornado	1

552 rows × 1 columns



Fire Dataset Integration

Kaggle'daki Türkiye orman yangınları veri setinin 2010-2015 yılları arasındaki verileri kullanılarak, kendi verimize hangi günlerde hangi şehirlerde yangın olduğunu Fire sütunu olarak ekledik.

Fire & No Fire

In [140]:
```
df.head(1)
```

Out[140]:

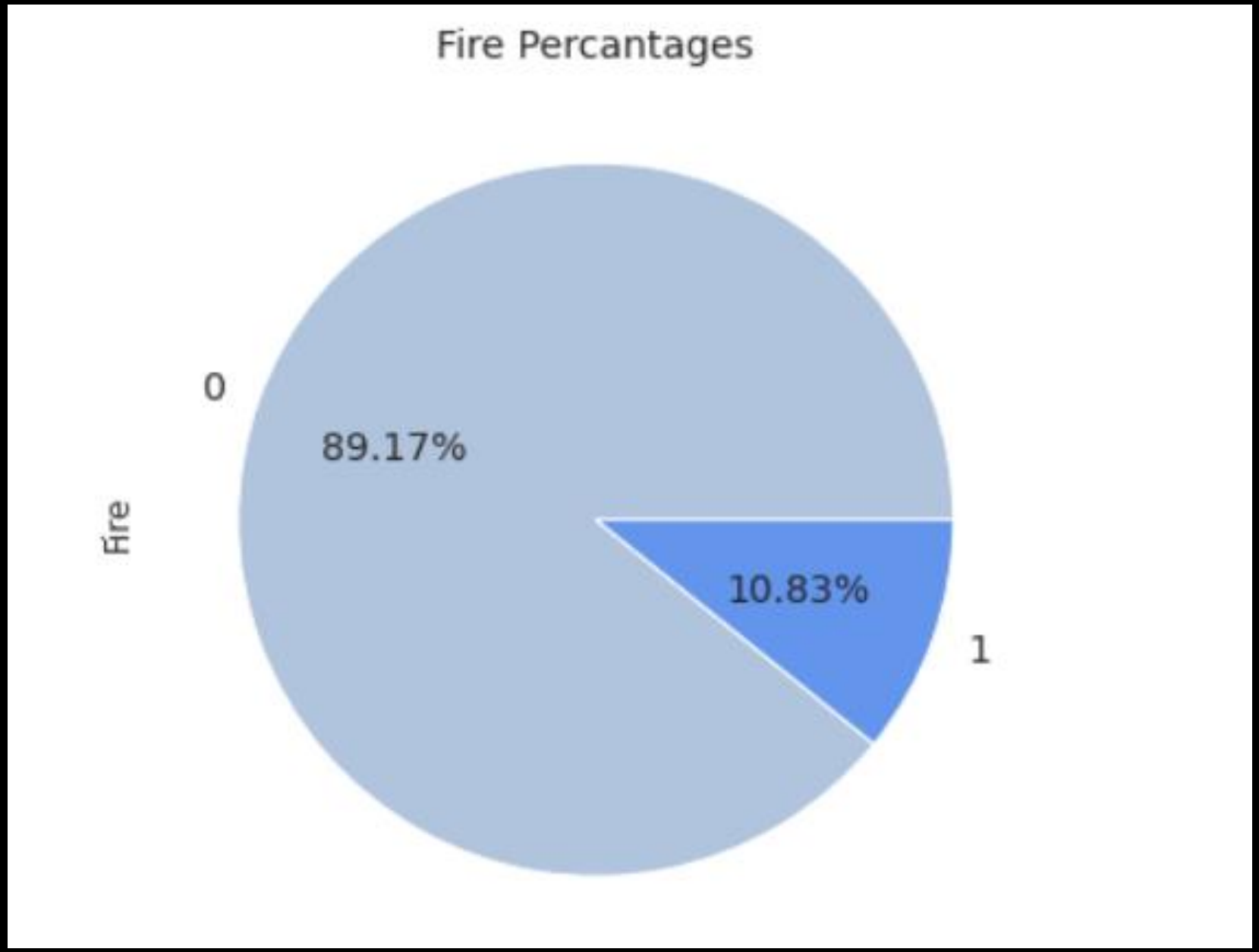
	City	Year	Month	Day	AvgDWS	Visibility	MaxSusWS	MaxWG	WeatherType	Date	TempAvgCel	TempMaxCel	TempMinCel	PrecipitationCm	DewPntCe
0	Adana	2010	01	01	3.8	6.0	10.24	29.92	Rain/Drizzle, Thunder	2010-01-01	13.9	19.0	7.0	0.0	8.

In [141]:
```
df_2010_2015 = pd.read_csv("df_2010_2015.csv")
```

In [142]:
```
df_2010_2015.head(1)
```

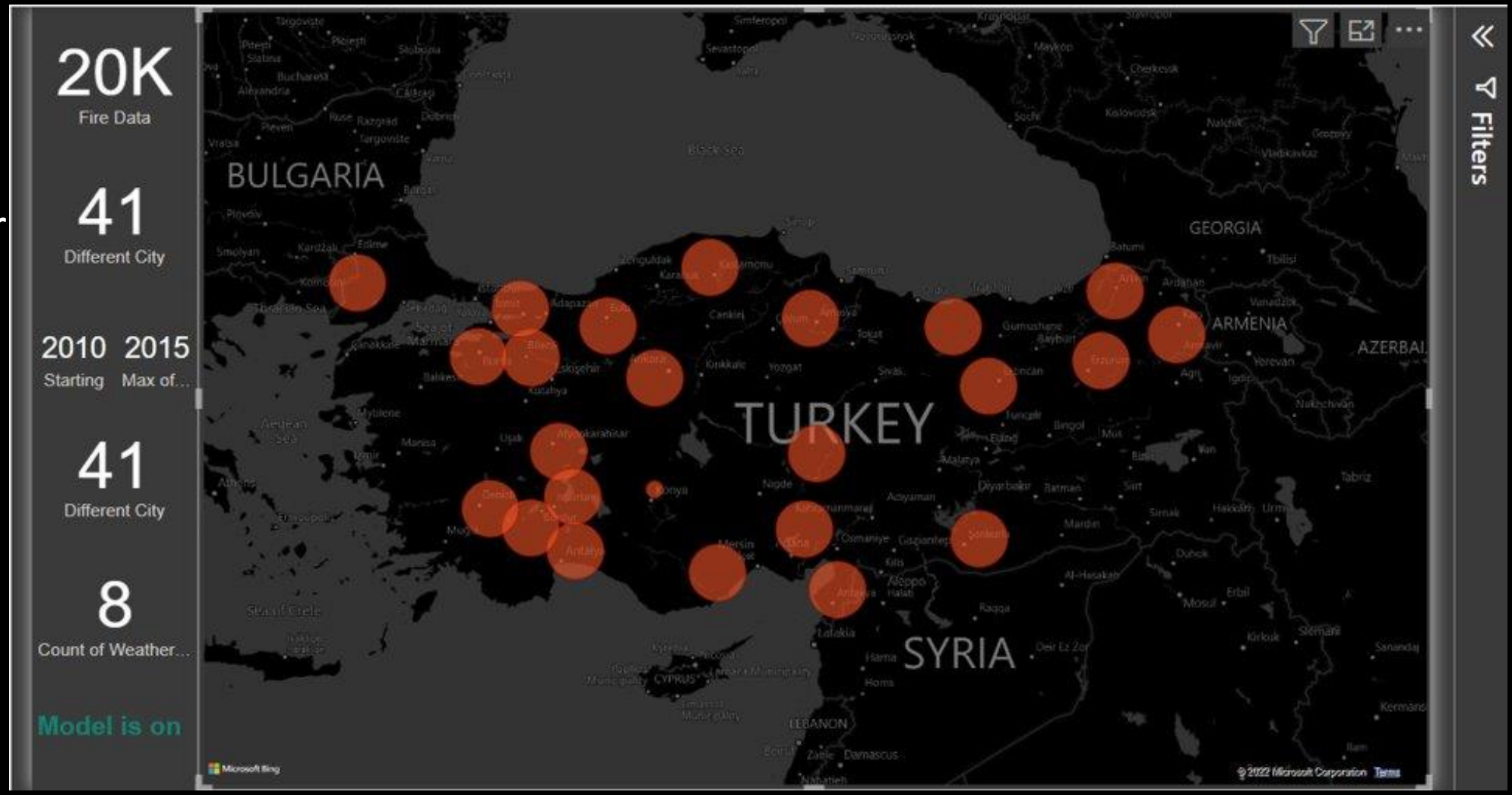
Out[142]:

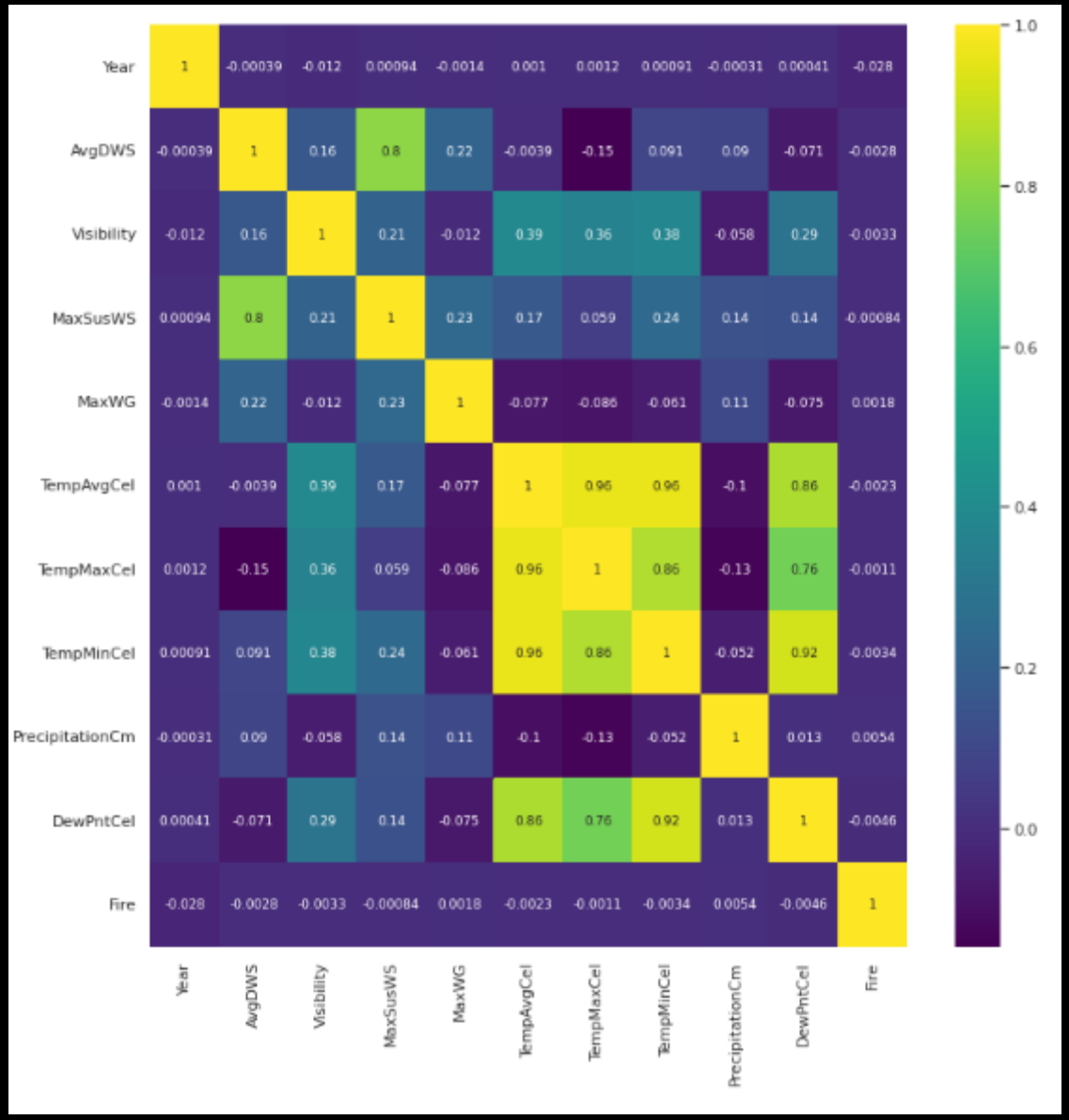
	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight	type	province	e
0	39.1908	31.1941	307.8	1.6	1.2	2010-01-01	915	Terra	MODIS	66	6.2	286.1	17.7	D	0	Afyonkarahisar	


Fire Percantages
Fire
0
89.17%
10.83%
1

Genel Bakış

- 20k rows
  - no fire -> 3652
  - Fire -> 3652
- 69 şehir -> 41 şehir




Year 1 -0.00039 -0.012 0.00094 -0.0014 0.001 0.0012 0.00091 -0.00031 0.00041 -0.028
AvgDWS -0.00039 1 0.16 0.8 0.22 -0.0039 -0.15 0.091 0.09 -0.071 -0.0028
Visibility -0.012 0.16 1 0.21 -0.012 0.39 0.36 0.38 -0.058 0.29 -0.0033
MaxSusWS 0.00094 0.8 0.21 1 0.23 0.17 0.059 0.24 0.14 0.14 -0.00084
MaxWG -0.0014 0.22 -0.012 0.23 1 -0.077 -0.086 -0.061 0.11 -0.075 0.0018
TempAvgCel 0.001 -0.0039 0.39 0.17 -0.077 1 0.96 0.96 -0.1 0.86 -0.0023
TempMaxCel 0.0012 -0.15 0.36 0.059 -0.086 0.96 1 0.86 -0.13 0.76 -0.0011
TempMinCel 0.00091 0.091 0.38 0.24 -0.061 0.96 0.86 1 -0.052 0.92 -0.0034
PrecipitationCm -0.00031 0.09 -0.058 0.14 0.11 -0.1 -0.13 -0.052 1 0.013 0.0054
DewPntCel 0.00041 -0.071 0.29 0.14 -0.075 0.86 0.76 0.92 0.013 1 -0.0046
Fire -0.028 -0.0028 -0.0033 -0.00084 0.0018 -0.0023 -0.0011 -0.0034 0.0054 -0.0046 1
Year
AvgDWS
Visibility
MaxSusWS
MaxWG
TempAvgCel
TempMaxCel
TempMinCel
PrecipitationCm
DewPntCel
Fire
1.0
0.8
0.6
0.4
0.2
0.0

Models

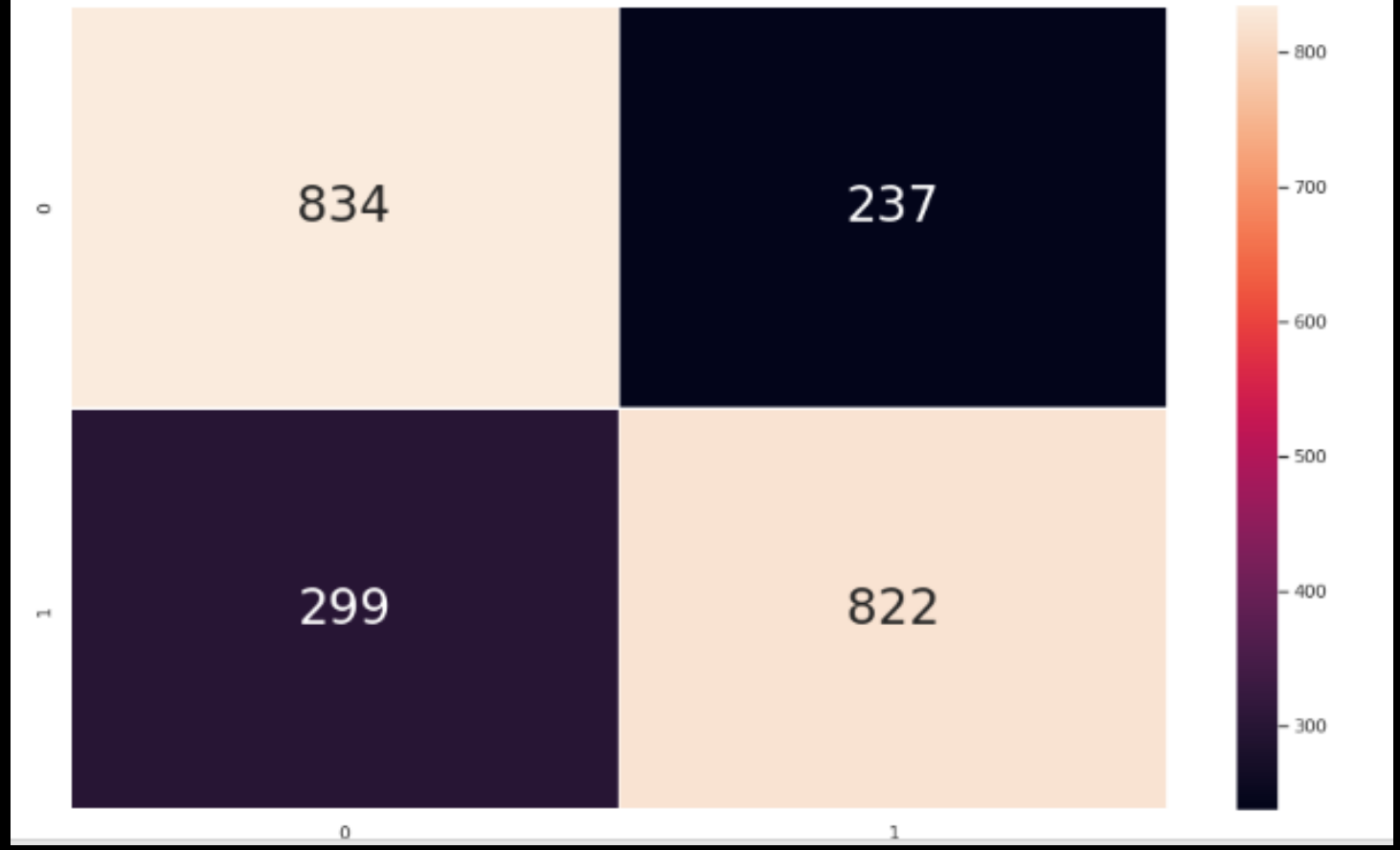
- Supervised classification
  - Logistic Regression
  - KNN
  - XGBoost
  - SVM
  - MLPClassifier

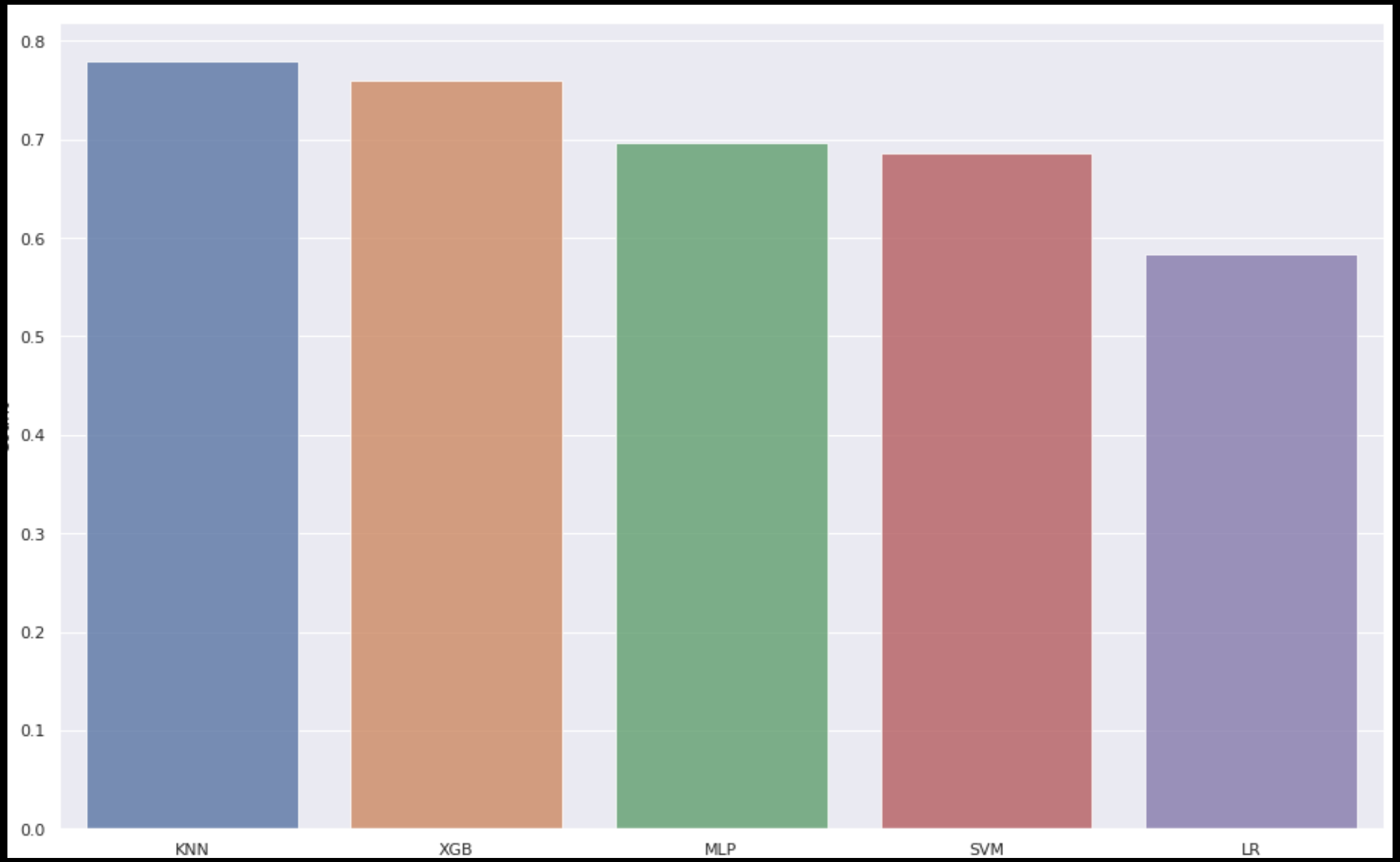
```
Training accuracy: 0.76
Test accuracy: 0.755
              precision    recall  f1-score   support

           0       0.74      0.78      0.76      1071
           1       0.78      0.73      0.75      1121

    accuracy                           0.76      2192
   macro avg       0.76      0.76      0.76      2192
weighted avg       0.76      0.76      0.76      2192

<matplotlib.axes._subplots.AxesSubplot at 0x7f9e75afccd0>
```




0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
KNN
XGB
MLP
SVM
LR