# Exploritory Analysis III

Exploritory Analysis of the Datasets to be used for
Machine Learning Predictions of Exoplanet candidates

The analyis will look briefly at:

- The Coandidates datset being used.
- The main features of the dataset
- If any of the features will be needed for the Classification Machine Learning

In [1]:

```
#imports
import pandas as pd
import numpy as np
import lightkurve as lk
import matplotlib.pyplot as plt
import seaborn as sn
```

In [2]:

```
# file import and dataframe creation
file = ('https://exoplanetarchive.ipac.caltech.edu/cgi-bin/nstedAPI/nph-nst
koi_df = pd.read_csv(file, low_memory=False)
koi_df.head()
```

Out[2]:

| | kepid | kepoi_name | kepler_name | koi_disposition | koi_pdisposition | koi_score | koi_fpflag_nt |
|---|---|---|---|---|---|---|---|
| **0** | 10797460 | K00752.01 | Kepler-227 b | CONFIRMED | CANDIDATE | 1.000 | 0 |
| **1** | 10797460 | K00752.02 | Kepler-227 c | CONFIRMED | CANDIDATE | 0.969 | 0 |
| **2** | 10811496 | K00753.01 | NaN | CANDIDATE | CANDIDATE | 0.000 | 0 |
| **3** | 10848459 | K00754.01 | NaN | FALSE POSITIVE | FALSE POSITIVE | 0.000 | 0 |
| **4** | 10854555 | K00755.01 | Kepler-664 b | CONFIRMED | CANDIDATE | 1.000 | 0 |

5 rows × 50 columns

In [3]:

```
# the shappe and size of the imported dataset
koi_df.shape
```

Out[3]:

```
(9564, 50)
```

These columns of interest can be used to check for possible data inbalance for the machine learning model as well as basic statistics in reagards to what will be outputted for the model classifications

**kepid** : int
arget identification number, as listed in the Kepler Input Catalog (KIC).
The KIC was derived from a ground-based imaging survey of the Kepler field
conducted prior to launch.

**kepler_name** : char
Kepler number name in the form "Kepler-N," plus a lower-case letter,

**koi_score** : float
A value between 0 and 1 that indicates the confidence in the KOI disposition.
For CANDIDATEs, a higher value indicates more confidence in its disposition,
while for FALSE POSITIVEs, a higher value indicates less confidence in that disposition.

**koi_disposition** : Char
The category of this KOI from the Exoplanet Archive. Current values are CANDIDATE,
FALSE POSITIVE, NOT DISPOSITIONED or CONFIRMED. All KOIs marked as CONFIRMED are also
listed in the Exoplanet Archive Confirmed Planet table

**koi_period** : double
The interval between consecutive planetary transits (days)

In [4]:

```
# apply a count of confirmed planets and false positives
confirmed_df = koi_df.copy()
confirmed_df = confirmed_df[confirmed_df['koi_disposition'].str.contains('C
                          confirmed_df['koi_disposition'].str.contains('FAI
confirmed_df = confirmed_df.groupby('koi_disposition')['kepid'].count()
confirmed_df = confirmed_df.to_frame()
confirmed_df.reset_index(inplace=True)
confirmed_df.rename(columns={'kepid':'count'}, inplace=True)
confirmed_df['%'] = (confirmed_df['count'] / confirmed_df['count'].sum())*1
confirmed_df = confirmed_df.sort_values(by='%', ascending=False)
confirmed_df.reset_index(drop= True, inplace=True)
confirmed_df
```

Out[4]:

| | koi_disposition | count | % |
|---|---|---|---|
| **0** | FALSE POSITIVE | 4840 | 64.498934 |
| **1** | CONFIRMED | 2664 | 35.501066 |

In [5]:

```
# plot a table
fig, ax = plt.subplots()

# hide the axis
fig.patch.set_visible(False)
ax.axis('off')
ax.axis('tight')

#create the table
table = ax.table(cellText=confirmed_df.values, colLabels = confirmed_df.col

# disply and save the table
name = "Candidate Disposition %"
plt.title(name, y=1.0, pad=-80)
fig.tight_layout()
plt.savefig('./graphs/candidate_disposition_percent.jpg', bbox_inches="tigh
plt.show()
```

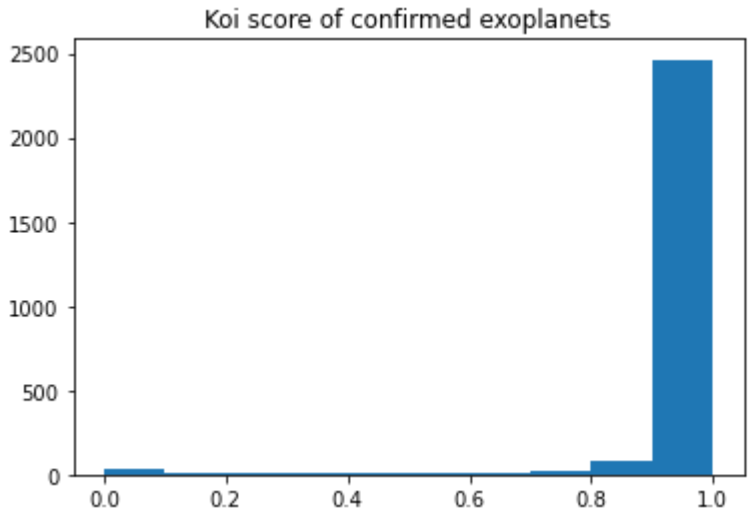
Candidate Disposition %

| koi_disposition | count | % |
|---|---|---|
| FALSE POSITIVE | 4840 | 64.49893390191897 |
| CONFIRMED | 2664 | 35.501066098081026 |

In [6]:

```
# plot the distribution of confirmed planets koi score
confirmed_exo_df = koi_df.copy()
confirmed_exo_df = confirmed_exo_df[confirmed_exo_df['koi_disposition'].str
plt.subplot()
plt.hist(confirmed_exo_df['koi_score'])
plt.title('Koi score of confirmed exoplanets')


plt.savefig('./graphs/confirmed_koi_score.jpg', bbox_inches="tight", dpi=45
plt.show()
```




In [7]:

```
confirmed_exo_stats = confirmed_exo_df['koi_score'].describe()
confirmed_exo_stats = confirmed_exo_stats.to_frame()
confirmed_exo_stats.reset_index(inplace=True)
confirmed_exo_stats.rename(columns={'index':'Stat', 'koi_score':'Value'}, i
confirmed_exo_stats
```

Out[7]:

| | Stat | Value |
|---|---|---|
| **0** | count | 2650.000000 |
| **1** | mean | 0.964119 |
| **2** | std | 0.137348 |
| **3** | min | 0.000000 |
| **4** | 25% | 0.992000 |
| **5** | 50% | 1.000000 |
| **6** | 75% | 1.000000 |
| **7** | max | 1.000000 |

In [8]:

```
# plot a table
fig, ax = plt.subplots()

# hide the axis
fig.patch.set_visible(False)
ax.axis('off')
ax.axis('tight')

#create the table
table = ax.table(cellText=confirmed_exo_stats.values, colLabels = confirmec

# disply and save the table
name = "Koi Score Stats (Confirmed)"
plt.title(name, y=1.0, pad=-60)
fig.tight_layout()
plt.savefig('./graphs/koi_score_confirmed_stats.jpg', bbox_inches="tight",
plt.show()
```

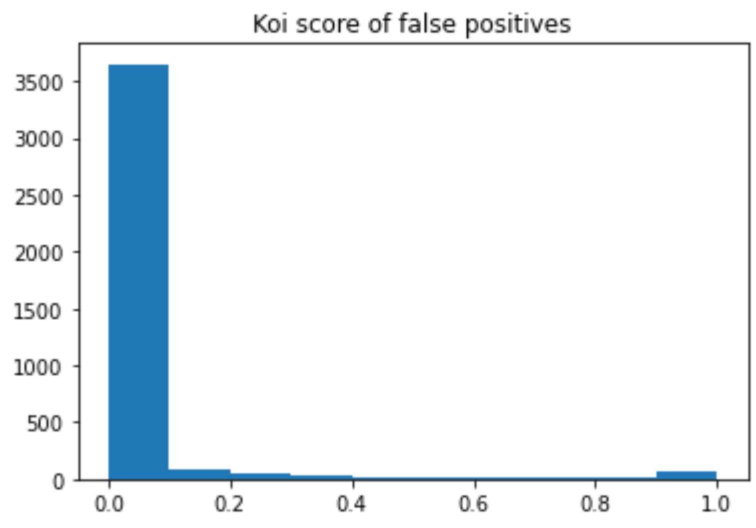
Koi Score Stats (Confirmed)

| Stat | Value |
|---|---|
| count | 2650.0 |
| mean | 0.9641192452830191 |
| std | 0.13734826532066208 |
| min | 0.0 |
| 25% | 0.992 |
| 50% | 1.0 |
| 75% | 1.0 |
| max | 1.0 |

In [9]:

```
# plot the distribution of false positive planets koi score
false_exo_df = koi_df.copy()
false_exo_df = false_exo_df[false_exo_df['koi_disposition'].str.contains('F
plt.subplot()
plt.hist(false_exo_df['koi_score'])
plt.title('Koi score of false positives')


plt.savefig('./graphs/false_koi_score.jpg', bbox_inches="tight", dpi=450)
plt.show()
```




In [10]:

```
false_exo_stats = false_exo_df['koi_score'].describe()
false_exo_stats = false_exo_stats.to_frame()
false_exo_stats.reset_index(inplace=True)
false_exo_stats.rename(columns={'index':'Stat', 'koi_score':'Value'}, inpla
false_exo_stats
```

Out[10]:

| | Stat | Value |
|---|---|---|
| 0 | count | 3946.000000 |
| 1 | mean | 0.038105 |
| 2 | std | 0.158799 |
| 3 | min | 0.000000 |

| | Stat | Value |
|---|---|---|
| **4** | 25% | 0.000000 |
| **5** | 50% | 0.000000 |
| **6** | 75% | 0.000000 |

In [11]:

```
# plot a table
fig, ax = plt.subplots()

# hide the axis
fig.patch.set_visible(False)
ax.axis('off')
ax.axis('tight')

#create the table
table = ax.table(cellText=false_exo_stats.values, colLabels = false_exo_sta

# disply and save the table
name = "Koi Score Stats (False Positive)"
plt.title(name, y=1.0, pad=-60)
fig.tight_layout()
plt.savefig('./graphs/koi_score_false_positive_stats.jpg', bbox_inches="tig
plt.show()
```

Koi Score Stats (False Positive)

| Stat | Value |
|---|---|
| count | 3946.0 |
| mean | 0.03810466294982267 |
| std | 0.15879922180278527 |
| min | 0.0 |
| 25% | 0.0 |
| 50% | 0.0 |
| 75% | 0.0 |
| max | 1.0 |

In [ ]: