Chris Johanson
Milestone Report for Capstone 1 Project

The data set being used for this project consists of astronomical radio data collected by the Parkes High Time Resolution Universe Legacy Survey. The survey's goal included searching for pulsar stars. A pulsar star is a type of neutron star that spins rapidly and produces radio energy emissions in the form of beams which are ejected from its magnetic poles. Pulsars also rotate at a somewhat regular interval, so the combination of the emissions and rotation patterns are detectable via radio telescopes here on Earth.

Although it is easy to think of space as being "empty," there is much more happening than the human eye can detect. Regarding signals of radio emissions, the Parkes High Time Resolution Universe Legacy Survey scanned across a certain band of radio frequency. Pulsars have slightly different emission patterns which are unique to the individual star, similar to the way all humans have fingers but the fingerprints uniquely identify an individual. If these specific emission patterns were the only detectable patterns, finding pulsars would not be so difficult. However, since there are a large number of diverse events occuring in space at any given time, especially when you consider all of the bands of the electromagnetic spectrum, it is extremely difficult to find the proverbial needle in the haystack. This is where machine learning plays an important role, particularly classification algorithms. The computational power of machine learning algorithms enable researchers to comb through large amounts of data with a reasonable degree of accuracy.

The implications of pulsar classification algorithms impact multiple fields within astronomy. Information about the nature and behavior of pulsar stars sheds light on topics such as: the density of matter (pulsars are second to black holes in terms of being the most dense objects we have detected thus far in the universe), Einstein's general relativity, and gravitational waves generated by supermassive black holes merging in the early universe.

The dataset that I am working with was downloaded from kaggle. It is a relatively straightforward dataset consisting of 8 features and 1 target column. The 8 features are the following:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.
9. Target class

Features 1 through 4 are statistical measurements of the integrated profile. As explained above, pulsars produce different emission patterns. This emission pattern varies with each rotation and is normally too weak to detect on its own. An integrated profile is the combining, or folding, of these weaker signals in order to increase the signal quality. Ultimately, the integrated profile is

what gives a pulsar star its unique identity and the increased signal quality makes detection easier. Features 5 through 8 are statistical measurements of the dispersion measure-signal to noise ratio, or DM-SNR curve. Dispersion measure represents the disruptive effect that the interstellar medium (matter/particles that exist between the star systems in a galaxy) has on signals generated by pulsars. The signal to noise ratio is the "strength of the signal compared to random noise" (4. Observations of Pulsars). When combined, the dispersion measure and signal to noise ratio increase the ability to make the necessary measurements when searching for pulsars.

A brief overview of the data demonstrated that there are, as stated above, 9 rows in total and 17,898 rows. Each row is, of course, an individual observation. To find out how many pulsars and non-pulsars there are in the dataset, the data was divided into those two groups using the target class column. Of the 17,898 rows, 16,529 (~90.8%) are not pulsars and 1,639 (~0.09) are pulsars. In terms of data types, the target class consists of an int64 while the remaining are float64. In addition, there are 0 null values.

In terms of cleaning and wrangling the data, the process was relatively simple and painless. The only issue that was encountered was with the columns names of the 8 features. There was a white space at the beginning of the column names and was removed with the following code:

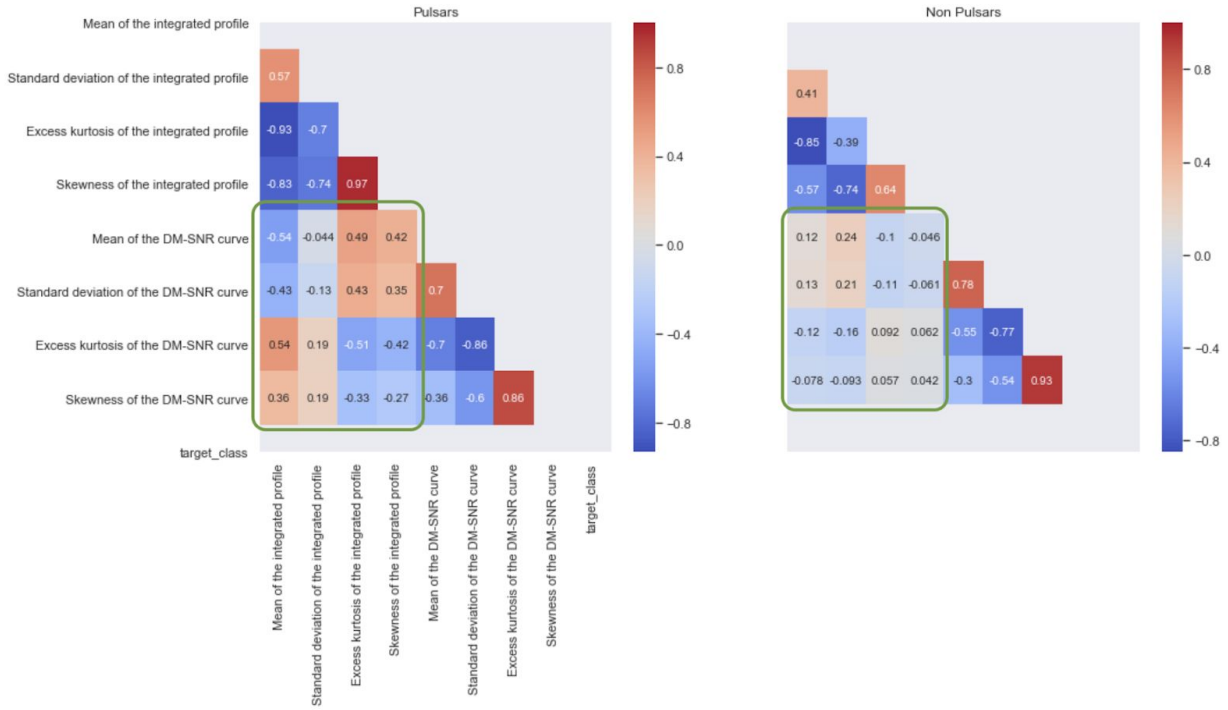
```
#Fix the white space in front of all the column names (except for target_class)
clean_column_names = []

for name in df.columns:
    clean_name = name.lstrip()
    clean_column_names.append(clean_name)

df.columns = clean_column_names
df.columns
```

```
Index(['Mean of the integrated profile',
       'Standard deviation of the integrated profile',
       'Excess kurtosis of the integrated profile',
       'Skewness of the integrated profile', 'Mean of the DM-SNR curve',
       'Standard deviation of the DM-SNR curve',
       'Excess kurtosis of the DM-SNR curve', 'Skewness of the DM-SNR curve',
       'target_class'],
      dtype='object')
```

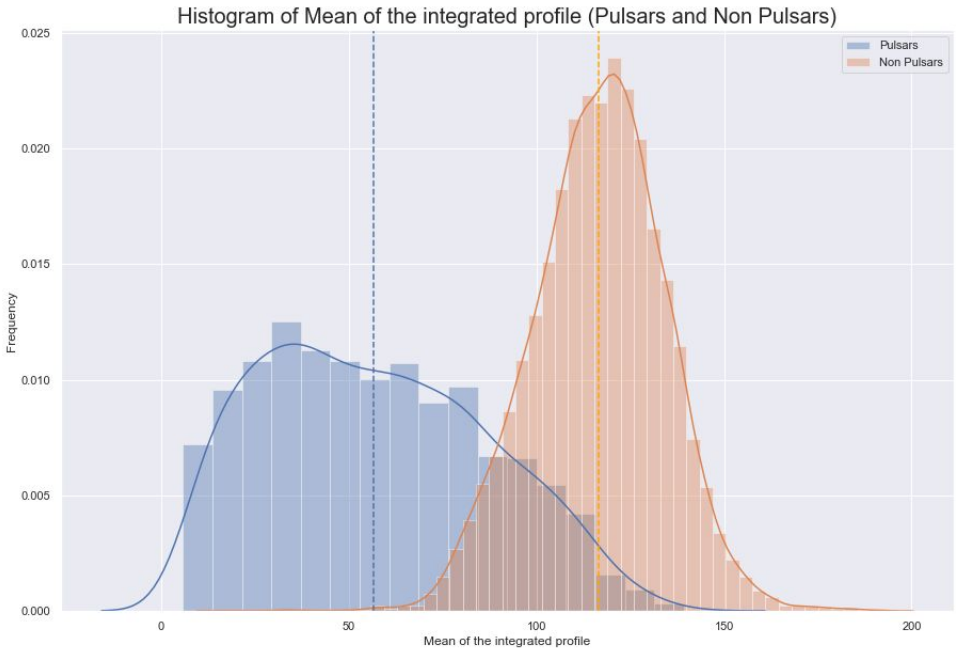
The exploratory analysis was straightforward. Being a classification problem, the first step was to take the entire dataset and split it in two according to the target class feature, where a "0" indicates that an observation is not a pulsar star and a "1" indicates that an observation is a pulsar star. Next, a correlation matrix heatmap was generated for both groups and plotted next to each other to investigate the relationships between each pair of variables:



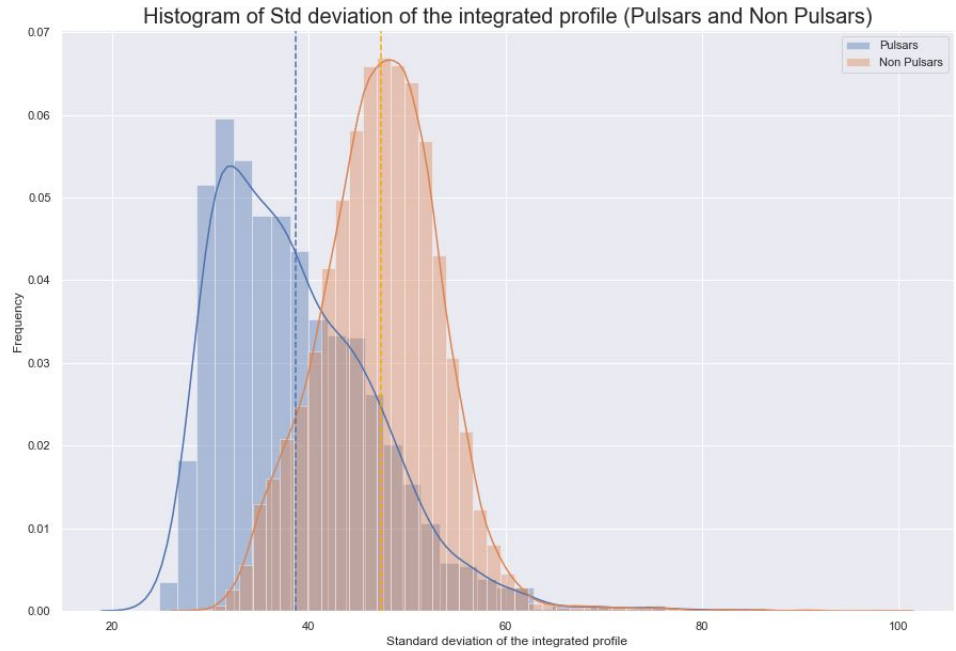
Comparing “Non Pulsars” and “Pulsars”, there are some interesting patterns. First, there are some relationships between variables that do not change, mostly occurring outside of the green rectangle seen in the visualization above. A lack of change is not surprising because the relative stagnancy of these elements is likely due to the fact that said elements either measure the integrated profile against itself or the DM-SNR curve against itself, as opposed to measuring an aspect of the integrated profile against the DM-SNR curve (or vice versa). Second, there are indeed noteworthy changes among the elements that fall within the highlighted green box. That is to say, elements comparing some aspect of the integrated profile and DM-SNR curve to each other. A number of these elements are rather neutral in the “Non Pulsars” group but become more polarized in the “Pulsars” group, for example the excess kurtosis of the DM-SNR curve with the mean of the integrated profile and the mean of the DM-SNR curve with the excess kurtosis of the integrated profile. It is too early to say as of right now, but these patterns merit further investigation when compiling the classification algorithm.

The exploratory analysis was concluded by dividing all 8 features into two groups: integrated profile and DM-SNR curve. The four statistical measurements (mean, standard deviation, excess kurtosis, and skewness) were compared between the two groups so the mean of the integrated profile could be compared with the mean of the DM-SNR curve, and so on. For all of the individual features, a dual histogram was plotted that included both cases of the two groups. Then, a two sample t-test was performed to make sure the result was statistically significant.

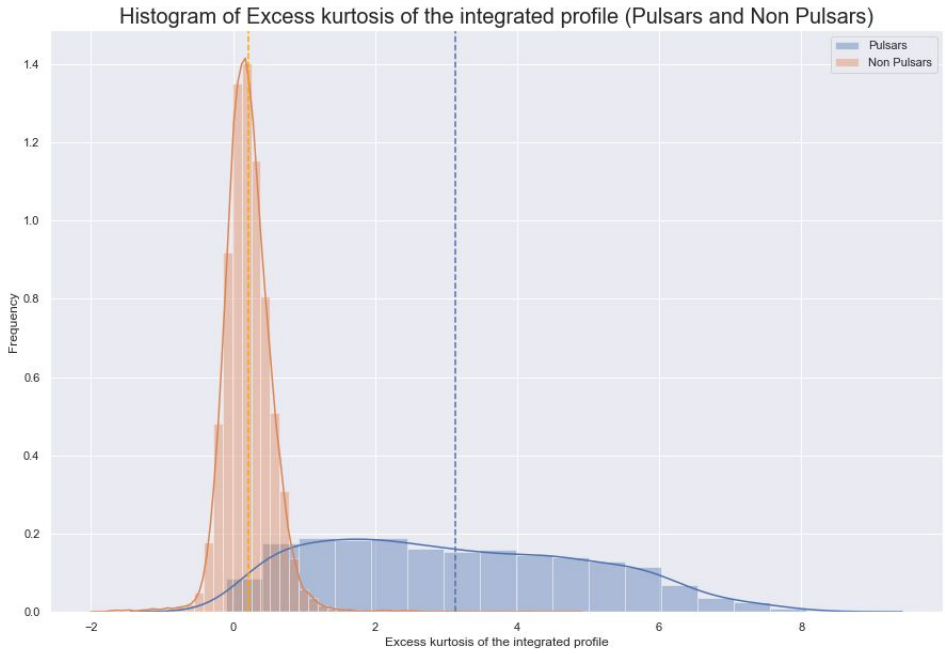
Starting with the mean of the integrated profile, the following dual histogram was generated:



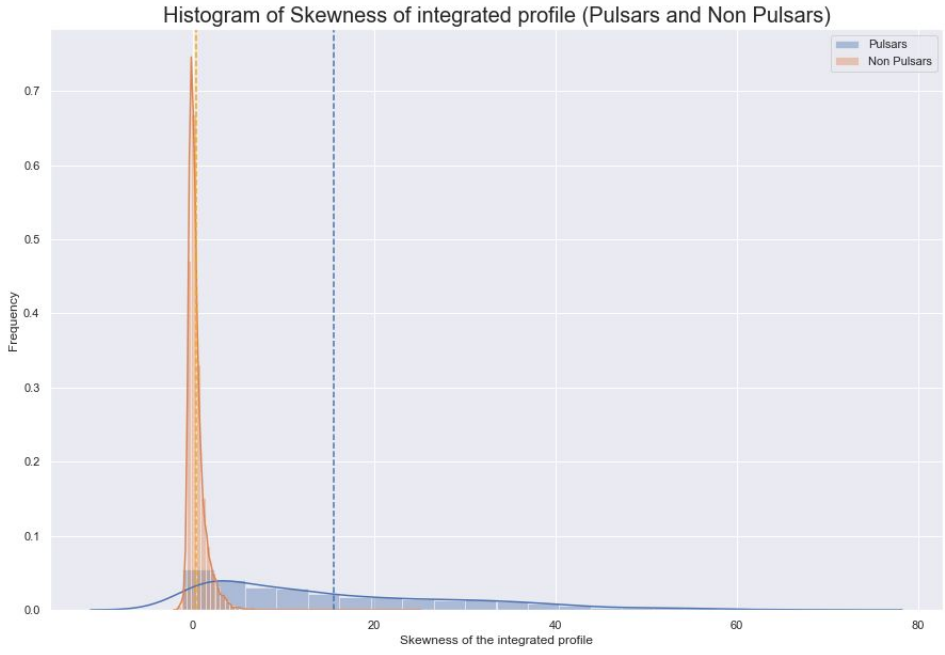
The two sample t-test for the mean of the integrated profile returned a p-value of 0.0, which statistically confirms that the two independent samples have different expected values. Next is the standard deviation of the integrated profile with the following dual histogram:



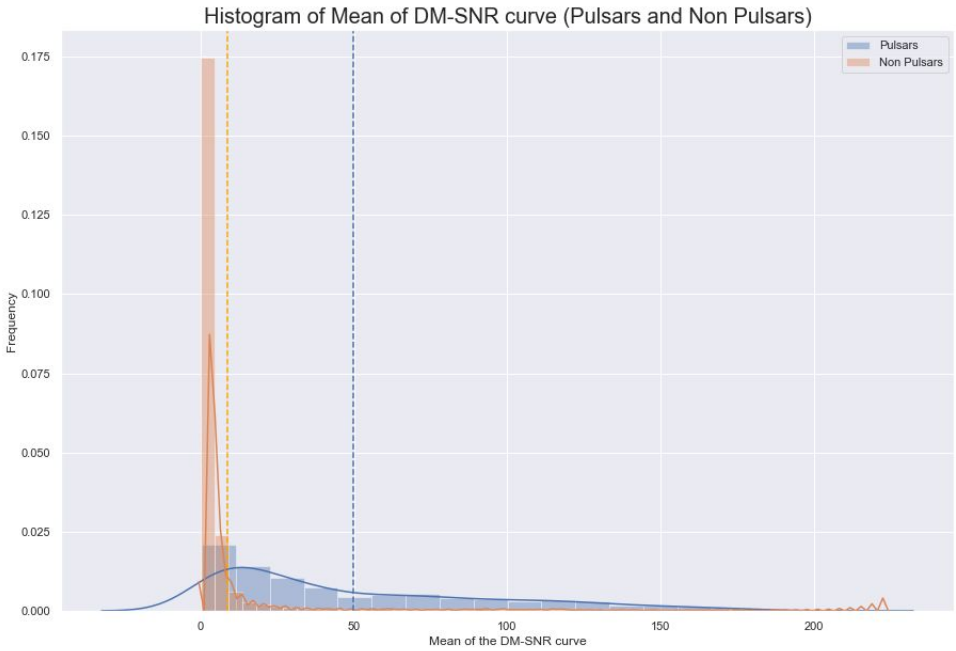
The corresponding two sample t-test for the standard deviation of the integrated profile returned a p-value of 3.99272404103977e-273, statistically confirming that the two independent samples have different expected values. For the excess kurtosis of the integrated profile, the following dual histogram was generated:



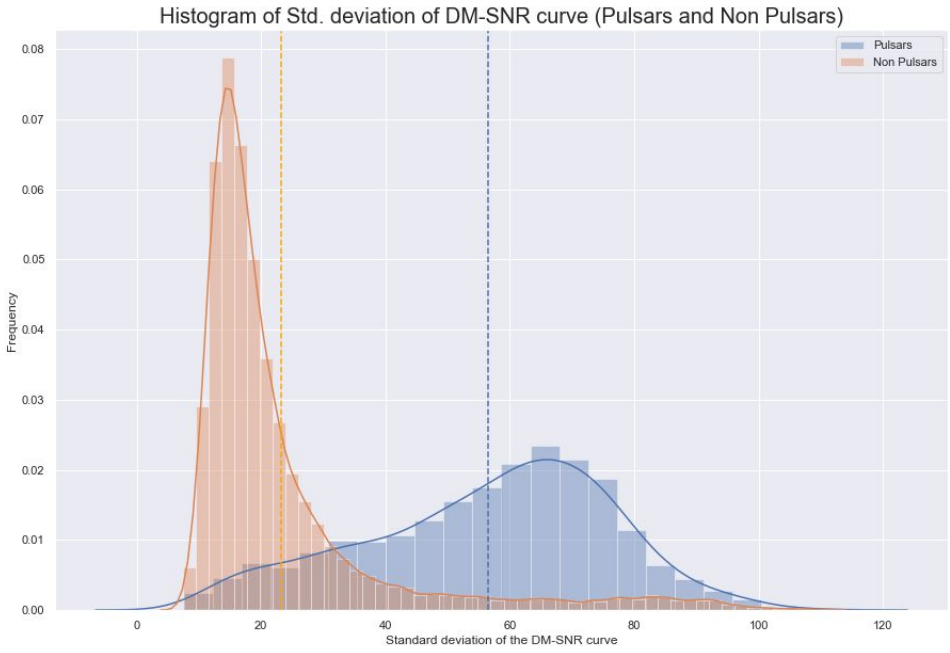
The two sample t-test for the excess kurtosis of the integrated profile resulted in a p-value of 0, which statistically confirms that the two independent samples have different expected values. For the skewness of the integrated profile, the following dual histogram was generated:



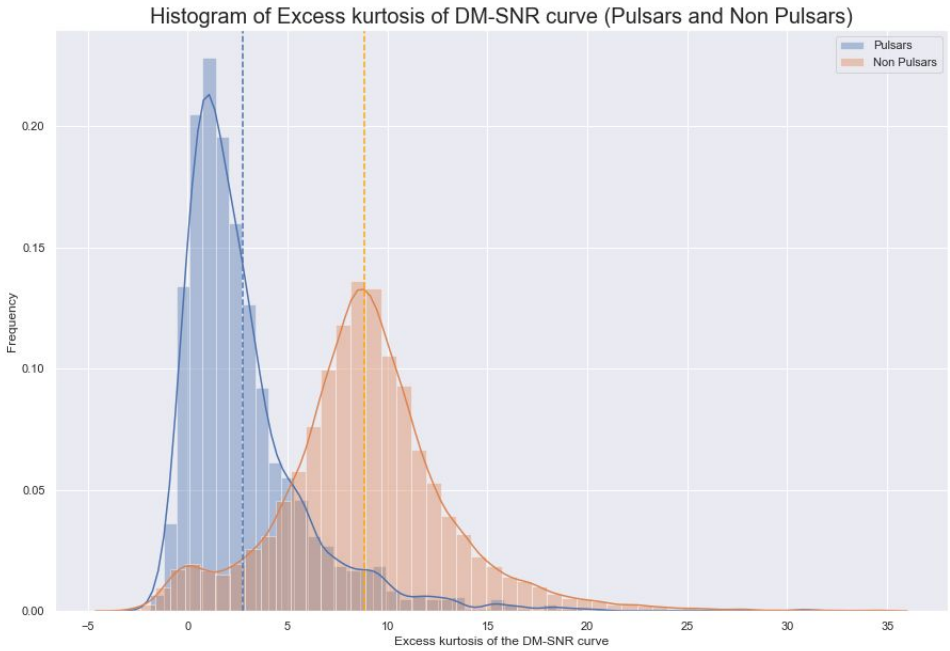
The two sample t-test for the skewness of the integrated profile returned a p-value of 8.735371914187084e-279, statistically confirming that the two independent samples have a different expected value. For the mean of the DM-SNR curve, the following dual histogram was generated:



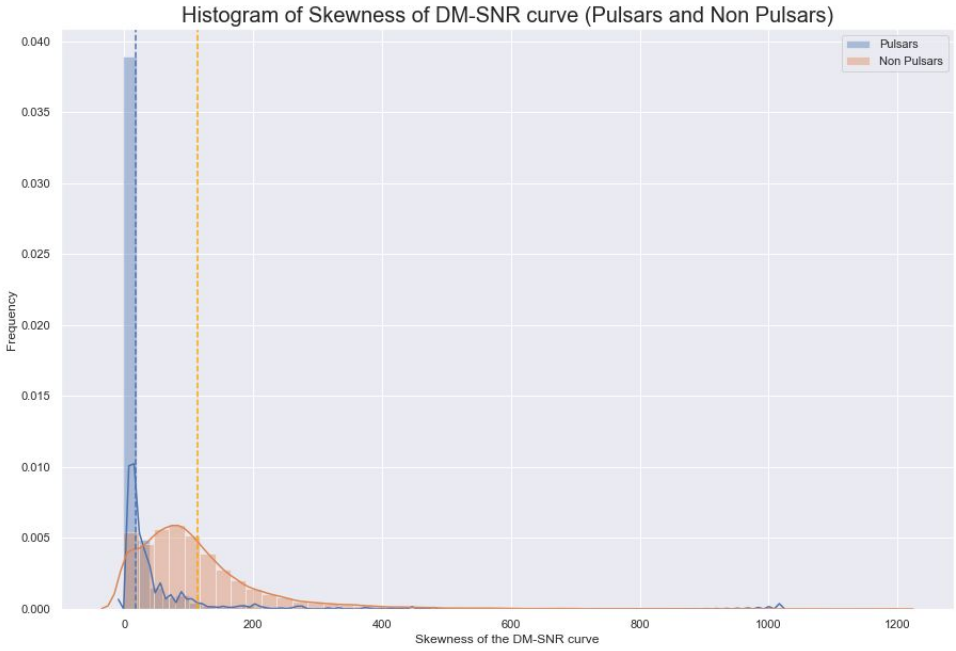
The two sample t-test for the mean of the DM-SNR curve returned a p-value of 2.7586700453147315e-213, which statistically confirms that the two independent samples have a different expected value. For the standard deviation of the DM-SNR curve, the follow dual histogram was generated:



The two sample t-test for the standard deviation of the DM-SNR curve returned a p-value of 0, statistically confirming that the two independent samples have different expected values. For the excess kurtosis of the DM-SNR curve, the follow dual histogram was generated:



The two sample t-test for the excess kurtosis of the DM-SNR curve returned a p-value of 0, which statistically confirms that the two independent samples have different expected values. For the skewness of the DM-SNR curve, the following dual histogram was generated:



The two sample t-test for the skewness of the DM-SNR curve returned a p-value of 0, statistically confirming that the two independent samples have different expected values.

In summary, pulsar stars play an important role within the field of astronomy and are particularly dynamic as a subject to research because of the impact they have on a variety of astronomical research topics. The findings of the exploratory analysis have brought forward a number of insights. First, the correlation matrix heatmap aided in highlighting interesting relations, with the most changes in correlation occurring when an integrated profile feature is compared to a DM-SNR feature. Said correlations will be strongly considered when a classification algorithm is constructed. Second, when the four statistical measurements (mean, standard deviation, excess kurtosis, and skewness) were compared between the integrated profile and the DM-SNR, the two sample t-test confirmed that the differences were statistically significant. The insights gained from the crucial exploratory analysis will play an meaningful role in constructing a classification algorithm that can be used to determine whether or not observations are a pulsar star.

References

1. "4. Observations of Pulsars." *4: Observations of Pulsars*, The University of Manchester Jodrell Bank Observatory, www.jb.man.ac.uk/distance/frontiers/pulsars/section4.html.