# \*\*Review of tips dataset\*\*

## \*\*Table of contents\*\*

\* [Introduction](#Introduction)

\* [Setup](#Setup)

\* [What is the tips dataset?](#What-is-the-tips-dataset?)

\* [Some statistics on the dataset](#Some-statistics-on-the-dataset)

\* [Description 1](#Description-1)

\* [Skewness and kurtosis 1](#Skewness-and-kutrosis-1)

\* [Covariance](#Covariance)

\* [Correlation](#Correlation)

\* [Plotting statistics](#Plotting-statistics)

\* [Description discussion](#Description-discussion)

\* [Regression analysis: Table bill vs tips](#Regression-analysis:-Table-bill-vs-tips)

\* [Best fit and mean fit lines](#Best-fit-and-mean-fit-lines)

\* [Regression discussion](#Regression-discussion)

\* [Some other analysis](#Some-other-analysis)

\* [Time of dinner](#time-of-dinner)

\* [Smoker](#Smoker)

\* [Description 2](#Description-2)

\* [Covariance and correlation](#Covariance-and-correlation)

\* [Skewness and kurtosis 2](#Skewness-and-kurtosis-2)

\* [Plotting other analysis](#Plotting-other-analysis)

\* [Discussion of other analysis](#Discussion-of-other-analysis)

\* [Further potential analysis](#Further-potential-analysis)

\*[Bibliography](#Bibliography)

## \*\*Introduction\*\*

In this notebook, it is intended to review the tips dataset, which can be found online. To do this, a number of approaches will be taken. The first step will be to review the data using standard descriptive statistics, such as mean, mode, median, quartile ranges, standard deviation, etc. Other information will also be gleaned, such as the skewness, and kurtosis of the data (this will be explained in the relevant section), as well as correlation and covariance between certain data sets.

A second step will be to conduct some analysis between the fields of total-bill and tip, to determine a relationship. This analysis will also include both a best fit, and a mean fit lines. These lines can be used for various reasons, and will be discussed in the regression analysis section.

The third section, two subsets of the data will be extricated from the data, and compared to each other. The two subsets of the data chosen for this analysis are the time that the diner at the location, as well as whether the diner was a smoker or not. Similar analysis will conducted with regards to these two data sets, as was conducted on the overall dataset. A discussion of the results can be found in the discussion of further analysis section.

A final section covers other potential analysis that could be undertaken, using the same dataset.

The tips dataset that was used in the below was sourced from [here](http://<https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv>). Note the data here is the raw form, to make it easier to import into our code.

\*Note: in this text, a number of ideas, or explanations may be mentioned. These have all been sourced either from the works listed in the bibliography, or from the author’s own ideas. For the sake of legibility, unless a direct quote, comparison, or specific suggestion is used, it will not be cited in the standard academic (Harvard) manner.

### Setup

Before running any analysis of the data, it is necessary to import the data, and the libraries that we will be using. In this case, it is intended to import the following libraries:

- \*\*Numpy\*\*: To run some functions for analysis.

- \*\*Pandas\*\*: The dataset will be held in various pandas dataframes. The pandas library will also be used to conduct various analytical exercises.

- \*\*Seaborn\*\*: Will be used for various plotting functionality.

- \*\*Matplotlib.pyplot\*\*: For use in some graphing functionality.

Additionally, throughout this notebook, the main dataframe that will hold the dataset will be called \*tips\_ds\*. The below code is used to import the libraries and dataset.

It should be noted that the above code does not have any outputs. The confirmation that the code was successfully imported can be found at the end of [What is the tips dataset?](#What-is-the-tips-dataset?) (below).

## \*\*What is the tips dataset?\*\*

The tips dataset is a set of data that is often used to teach analytical skills, especially around the segregation of data, and analysis of a group, or a whole set. Additionally, it could be used for extrapolation of data, where the data in question is not included in the actual dataset. The data forms part of a tutorial for the Python Seaborn package, and is used to demonstrate how to conduct plotting of data in Python.

The tips dataset was collected in a restaurant. (Jain, 2017) suggests that the data was collected over a 2.5 month period, in a branch of a chain restaurant in California, USA. The data was collected by a server (waiter/waitress), and recorded each table they served, and included a variety of information regarding the table.

This information that was recorded fell into 7 different columns. These were:

- \*\*total-bill\*\*: the total\_bill for each meal (numerical values to two decimal places).

- \*\*tip\*\*: the tip received for each meal (numerical values to two decimal places).

- \*\*sex\*\*: gender of the table (either male or female. Doesn’t note if this was the gender of the payer, but would be safe to assume so).

- \*\*smoker\*\*: whether the table being served was in a smoking or non-smoking section (this was California in the 1990s after all)

- \*\*day\*\*: day of the week that the meal was being served. Covers only Thursday to Sunday.

- \*\*time\*\*: the time of day that the service was being conducted. Broken down as either lunch or dinner service.

- \*\*size\*\*: the number of people in the party (whole numerical value).

In total the collector made 244 entries into the dataset.

All the above data can be confirmed by Python.

\*\*\*

## \*\*Some statistics on the dataset\*\*

In order to be able to work with the dataset, and draw any conclusions from the data, it is important to determine some of the dataset’s properties. In order to do this, we will extract basic measures, more commonly known as descriptive statistics. These statistics can then be used as a guide to further analysis, and comparison.

#### Description 1

The initial set of descriptive statistics are the mean, mode, and median of the data, as well as standard deviation, quartiles, and min and max values. Luckily, pandas has the ability to provide the values for most of these statistics, using the \*describe\* function. However, while this is useful, it is also necessary to understand what the terms provided by the \*describe\* function actually mean.

- \*\*Mean\*\*: Also known as the simple average, is the sum of all the values divided by the number of values being summed.

- \*\*Standard deviation\*\*: A measure of how are a number is from the mean. In a perfectly normal distribution, ~68% of all values would be within 1 standard deviation of the mean. Represented in the \*describe\* output as \*std\*.

- \*\*min\*\*: The lowest value within the dataset.

- \*\*25%\*\*: The value for the 25th percentile. In other words, 25% of all the values in the dataset are below this value.

- \*\*50%\*\*: The value for the 50% percentile. In other words, 50% of all the values in the dataset are below this value. This value is often called the median value.

- \*\*75%\*\*: The value for the 75th percentile. In other words, 75% of all the values in the dataset are below this value.

- \*\*Max\*\*: The largest value in the dataset.

The values for the tips dataset are retrieved as follows:

As we can see, certain measures from the descriptive statistics such as mode, skewness, and kurtosis are missing from the \*describe\* functionality. These statistics also give rise to important information regarding the dataset. The [Skewness and kurtosis](#Skewness-and-kutrosis-1) will be dealt with in their respective section. This leaves the mode to be ascertained.

The mode of a set of numbers, is the number that appears most often within that set. It is not necessarily close to the median or mean of a dataset, and can be well outside the 1st standard deviation of the data (this can happen when the distribution of the data is skewed). If no number appears most often (i.e. no number appears more often than any other number), then there is no mode for the data. Additionally, unlike the mean, and median statistics, which need to have numbers to work, the \*mode\* function can work on any set of data.

Using pandas, the mode can be determined using the \*mode\* function.

The \*describe\* function only returns data for numerical fields, while the \*mode\* function included information on the mode of the columns that contained string. The \*describe\* function can be used to provide some descriptive statistics on the data in the non-numerical columns. This is done by the inclusion of the \*include=[np.object]\* argument in the \*describe\* function. In addition to the statistics returned in the original extract, the following statistics are provided:

- \*\*Unique\*\*: The number of unique values to be found in the column.

- \*\*Top\*\*: The value returned most frequently in the column. Should return the same value as the \*mode\* function.

- \*\*Freq\*\*: The frequency of the most commonly returned value from top/mode.

Using this \*describe(include=[np.object])\* function, the following values are returned for the tips dataset.

#### Skewness and kurtosis 1

The skewness and kurtosis are interrelated terms that are used to describe the nature of the distribution of the dataset, and how it differs from a normally distributed dataset. The definition of these terms are:

- \*\*Skewness\*\*: the direction and amount of asymmetry of the dataset about its mean. If the absolute value of the skewness is:

- greater than 1, the data is highly skewed and the distribution is very asymmetric.

- greater than 0.5 and less than 1, the data is moderately skewed, and the distribution is somewhat asymmetric.

- greater than 0, and less than 0.5, the distribution of the data is approximately symmetric.

- equal to 0, the data is normally distributed, and symmetric.

The sign of the skewness (i.e. positive or negative) also determines the skewness. Negative skewness indicates that the distribution is skewed to the left, the mean being less than the median, which is less than the mode. Positive values of skewness indicate the opposite, with the distribution being skewed to the right, and the mode being less than the median, which is less than the mean.

- \*\*Kurtosis\*\*: the kurtosis of a dataset indicates the sharpness, or flatness, of the peak of the data (around the mode, or mean, depending on the skewness).

The kurtosis is measure against the normal distribution, which has a kurtosis of 0. If the kurtosis is negative, then distribution of the data has a smaller standard deviation, as more values are grouped near the mean. This gives the distribution a sharper, and higher peak, and narrower body. A positive kurtosis indicates that there is less grouping around the mean, and indicating that the distribution has a larger standard deviation. This also gives the distribution a flatter, and lower peak, and a wider body.

In pandas, the skewness and kurtosis of a dataset can be ascertained using the \*skew\* and \*kurt\* functions. These functions return values for each numeric column within the data set.

#### Correlation

Correlation is a statistic that can be used to measure how well two sets of data correspond to each other. Weisstein (2019) defines correlation as \*“the degree to which two or more quantities are linearly associated.\* As such, a correlation coefficient can be calculated that shows the relationship between the two sets of variables, as well strength of the relationship.

In correlation analysis, positive values show that the two sets of data are positively correlated (i.e. as one value rises or falls, so does the other). Conversely, negative values indicate that the two data sets are negatively or inversely correlated (i.e. as one value rises, or falls, the other falls, or rises). A zero value indicates that there is no relationship between the two sets of data. The strength of the relationship is indicated be the actual value of the correlation coefficient. An absolute value above 0.5 is considered a strong correlation, and above 0.75 is a very strong correlation. A value of -1 or 1 means that the two sets of data are perfectly correlated (i.e. either perfectly positive or perfectly negative correlation).

In pandas dataframes, the \*corr\* function can be used to ascertain the correlation between numeric sets of data.

As there are 3 sets of numeric data, the correlation gives a 3x3 grid of data (with the diagonal self-correlation coefficient blanked out).

#### Plotting statistics

Before a discussion of the statistics that were determined, it is useful to plot some of the statistics determined. In this case, it would be useful to plot the \*total\_bill\* and \*tip\* columns, which will show the skewness of the distribution.

#### Covariance

#### Correlation

#### Description discussion

The descriptive statistics for the \*tip\* dataset give a fair detail with regards to the data.

##### Correlation

If we work in reverse order of statistics provided, we can see the relationship between the \*total\_bill\* and tip values. The correlation value is 0.676, which is in the upper 1/3 of the strong correlation. It is also positive. This suggests that there is a strong relationship between the size of the bill and the tip. There is also a strong relationship between the \*total\_bill\*, and the size of the party. Both of these are to expected, as generally, the larger the party, the larger the bill.

There is, however, a medium to strong relationship between the size of the party, and the left. This is somewhat surprising, as \*tip\* in the USA are generally a fixed percentage of the bill. This could indicate that the larger the party, the less generous the tipper felt. However, it should be cautioned that a large party leaving a small tip could affect this outcome. The removal of the outliers for the party size (75% of parties were 3 people or less, while the largest was 6 people) could increase this correlation.

##### Skewness and Kurtosis

Looking first at the kurtosis, the value of 3.65 for the tip data set suggest that there will be a distribution with sharp peak, and narrow body, with few outliers, when plotted. By contrast, the kurtosis for the \*total\_bill\* is given as 1.22. This suggests a more humped distribution, with a low, smooth peak, and a wide body, with numerous returns in the tails to either side. The plot of both the \*total\_bill\* and tip values shows that the kurtosis expectations are met.

Visually, it can be seen that both plots are skewed to the right, with the mean greater than the median. This is as expected, and is confirmed by the returned values for mean, mode, and median, that were calculated in the first section of the statistics.

To note: it is somewhat difficult to distinguish between the mean and median for the \*tip\* data, as they are very close. However, when splitting out the \*tip\* data, and plotting separately, they can clearly be seen.

As the skewness for both sets of data in greater than 1 (by 13% and 47% respectively), this indicates that the data is the set is very asymmetric. This is clearly demonstrated by the plotted graphs.

Overall, the skewness corroborates the data gleaned from the initial descriptive statistics for the mean, median, and mode.

##### Standard deviation and quartiles

Once again looking at the \*tip\* and \*total\_bill\* sets of data, the initial descriptive statistics gave a standard deviation of 8.9 and 1.34 respectively. This would indicate that that ~68% of all the data points for these two data sets would fall into a range of the mean less the standard deviation to the mean + the standard deviation. This can be checked using the following script.

Based on the results above, we can see that for both the \*total-bill\* and \*tip\* sets of data, the % of results that fall into the first standard deviation either side of the mean is, respectively, ~4% and ~6.5% greater than expected. This indicates that something is either wrong with the standard deviation or the data in the dataset. However, Smith (2016) notes that the standard deviation should only be used in conjunction with the normal distribution, as using it with asymmetric data will give false results. As we know from the above plots and skewness calculations, the data in the data set is not normally distributed, and is highly asymmetric. This accounts for the larger than expected size of the data, when using the lower and upper bounds of standard deviation.

Smith (2016) suggests that instead of standard deviation for skewed distributions, it is better to examine the quartiles. The quartiles, unlike the standard deviation, show the values where the data set can be broken into 4 even segments (i.e. 1st, 2nd, 3rd, and 4th segments), so that each segment has the same number of results in it. As there are 244 results in the data set, this would suggest that the quartiles contain 61 (244/4) results in each section. Additionally, the range between the upper bound of the 1st and 3rd boundaries is called the interquartile range, and contains 50% of all the values in the set. While this may sound intuitive, it does allow a method to determine the 50% of the data set that are surrounding the median value (which is the upper bound of the 2nd quartile). This allows for a filter to be applied to the data, and can help reduce the impact of outliers on some of the statistics relating to the data.

Looking at a modified table of the initial statistics below, we can see how outliers can impact the median, mean, and mode of the data set.

As can be seen from both sets of data, the maximum value for is considerably larger than the median (50% quartile) value. In the \*table\_bill\* set of data, it the max value is nearly twice the median value plus the 75% quartile value. In the \*tip\* data set, the gap between the 75% quartile value and the max value is more than twice the median value. It is also worthwhile noting that the value for the interquartile range is 21.2% and 15.6% respectively of the max value. This confirms the data from the histograms in the plots above, which showed clustering around the median.

##### Other data

The non-numerical data has not been looked at closely, due to the fact that there is little manipulation that can be done with it. However, it does show some interesting results that may be useful for further analysis.

In looking at the returns for the non-numerical values, we can see that the number of records where men (157/244 = 64.3%), non-smoking (61.89%), dinner (72.1%) form the majority of entries in their respective sets of data. While eating on a Saturday only accounts for 35.66% of entries, it is the largest set of days recorded. This suggests that while not easily manipulable, in the same way as numerical data, they do present viable filters to be used in further analysis.

Additionally, the \*size\* of the party being served has not been examined closely, except for the correlation values. This is due, in part, to the small range of the values (1 – 6 people per record), and the fact that 25% quartile is the same as the median value, and 75% of the data is between 1 – 3 people. While its kurtosis is not as dramatic as the \*tip\* set of data, its skewness is almost the same. This suggests that the \*size\* set of data would have a reasonably smooth peak, similar to the \*table\_bill” set of data, but that it is almost as right skewed and asymmetrical as the \*tip\* data set.

##### Conclusion of descriptive statistics

While it is difficult to draw too many conclusions from the initial set of statistics, some answers can be drawn. The data, in both cases, is asymmetric, and positively skewed, as demonstrated by many of the statistics identified. Due to this, such data as the standard deviation is not useful, and should be disregarded. Additionally, in both cases, there are large values in the tails of the data set which influence the descriptive statistics of the data, such as the mean.

In order to gain some perspective on the actual information in the data, it may be necessary to remove any value outside of the median, +/- 1.5 x the interquartile range. Another approach would be to include the non-numerical columns as filters, so that each day, service session, gender, or smoking, to the dataset and compare the results. It may be possible to determine if there is a set of patrons that is generally more generous with the tips being given out, or if the gender, or smoking habits affects the tip percentage. These could all be determined in a later stage of this document.

## \*\*Regression analysis: Table bill vs tips\*\*

As noted above there is a clear correlation between the between the \*total\_bill\* and \*tip\* sets of data. However, a cursory glance at the top and bottom values for each data set shows that the relationship is not the same for each entry (i.e. the tip is not the same percentage of the bill in each case). This can be seen below

As we can see from the values above, for the top/bottom three values in each category, there are some overlaps. Row 170 has the highest \*total\_bill\* and \*tip\*, but does not have the highest \*percentage\_tip\*. Row 67 appears in all three sets information, being the lowest value in \*total\_bill\*, the joint lowest value in \*tip\*, but the third highest value in \*percentage\_tip\*. It should also be noted the rows with lowest values in for \*total\_bill\* are the same rows for lowest \*tip\*.

There is no clear and obvious relationship between the \*total\_bill\* and \*tip\*, using the highest and lowest values. Therefore, it may be possible to determine a relationship visually, using a graph. A scatter graph, with the values plotted against each other would give an indication of a relationship.

On visual inspection, it looks like there could be a general linear pattern, from the point [10, 1] toward the point [50, 8]. This line would seem to be reasonably tight between the 10 – 20 marks on the x-axis, but spreading out more widely from the 25 mark on the x-axis. However, this analysis would not be exact, and is only an impression based on visual observation.

#### Best fit lines

A better way to determine any relationship between the \*table\_bill\* and \*tip\* value would be to run a calculation that could calculate a line that would “fit” between all the values. IT is possible that we could take the highest and lowest points (as found above), and draw a line that connects them both). Alternatively, it may be possible to use the second or third highest and lowest points, and draw a line between them. However, these lines would all have different slopes, and different formulae. Surely they all can’t be right?

There are a number of methods of determining the best fit line. Seaborn and Numpy both have methods of calculating best fit lines. For Seaborn two methods are \*regplot()\* and \*lmplot()\*. \*regplot()\* is a simple linear regression model, while \*lmplot()\* combines \*regplot()\* and \*FaceGrid\* functionalities. Numpy uses \*polyfit()\* which is a \*”least squares polynomial”\* regression method (SciPy.org, 2019). As Waskom (2018) notes that \*regplot()\* and \*lmplot()\* return the same data, there is no need to plot them together. However, \*polyfit()\*, and \*regplot()” return slightly different values, as can be seem with the slight difference on the graph.

(Chubarov, 2017)

If the degrees of fit are changed on \*polytfit()\* from 1 to 2, and the \*ci\* value in \*replot()\* is switched on, more pronounced differences can be seen. (Note: this is only to show the differences in the functions).

#### Regression discussion

As we can see from the regression lines there is a distinct relationship between the values in the \*total\_bill\* and the \*tip\* sets of data. While the correlation analysis performed earlier shows the extent to which the tewo values move in conjunction, the regression analysis shows how the value of \*total\_bill\* impacts the value of \*tip\*.

The formula for a line is given as \*y = mx + c\*, where y is the expect value of the dependent variable, in our case \*tip\*; m is the slope of the line, as well as the modifier for the plotted variable (\*total\_bill\*), while is an independent modifier for the formula. Using the value derived in from the \*polyfit()\* function, we can determine the formula for the line of regression.

As we can see the slope of the line is 0.105, which is a positive number. This is confirmed by visual inspection of the graph. What this means is that there is a positive association between the \*total\_bill\* values, and the \*tip\* values, and as one increases, the other is expected to as well. Additionally, we can see the lines that were plotted earlier, using the smallest and largest points were correctly aligned, however, the slopes were off by about 40% for the closest line.

Regression analysis, especially in statistics, can used to determine unknown variables, based on a known variable. In this case, it should be possible to determine, using the formula for the regression line, what the value of a tip would be, if the value of the bill was known. This could be shown using random variables, and plotting them on the graph.

It should be noted that as the values in the \*total\_bill\* data set grew larger, their distance from the line of regression also increased. This would suggest that while the regression line is useful for calculating the tip for bills between 9.00 and 30.00, actual tip values outside this range might differ considerably from what has been calculated.

## \*\*Some other analysis\*\*

### Time of dinner

### Smoker

#### Description 2

#### Covariance and correlation

#### Skewness and kurtosis 2

#### Plotting other analysis

#### Discussion of other analysis

## \*\*Further potential analysis\*\*

## \*\*Bibliography\*\*

# Bibliography