# Descriptive Data Analysis Notes

Before starting with descriptive data analysis, we contacted the client to solve the previous doubt about the field Senior. They told us that the Senior category should include 65 year old subscribers. Therefore, we applied the following conditional function to that column:

=SI([@Age]>65;"Yes";"No")

Once we had completed data cleaning and transformation, we copied the excel file into a new folder called Data Analysis and then we renamed that file to “amazon\_churn\_descriptive\_analysis.xlsx”. Descriptive data analysis was performed on that file.

For this purpose, we used the Data Analysis tool that is already integrated in Excel, which enabled us to obtain a quick summary of the most relevant descriptive statistics: mean, median, mode, standard deviation, variance, minimum, maximum, range, symmetry coefficient and kurtosis coefficient.

## Descriptive analysis on numeric columns

* **Account Length (in months):** the average account length is approximately of 32 months. However, the majority of the subscribers are new clients as the mode is equal to 1 month. Data dispersion regarding account length is notorious: standard deviation is equal to 24 months, meaning that there are some extreme values (very new accounts and/or very old ones). This is confirmed by the range value of 76 months: there is a minimum of 1 month (the mode) and a maximum of 77 months (probably a very loyal client). Moreover, negative kurtosis shows that values are concentrated around the mean (lighter tails than the normal distribution). However, the standard deviation was big in comparison with the mean. These statistics reinforce the idea that there are some individuals with very atypical values (outliers). Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some atypically high values (right tail).
* **Customer Service Calls:** the average number of calls is less than 1. Moreover, the majority of the subscribers never called, as the mode is equal to 0. Same for the median. On the other hand, data is really dispersed: standard deviation is equal to 1.41 (1.5 times the mean) meaning that there are some extreme values, most probably customers that have called a lot of times, due to problems with the service or complaints versus clients not calling at all. This is confirmed by the range value of 5 calls: there are clients which have called a maximum of 5 times. The positive kurtosis shows that values are more dispersed from the mean and with stronger tails than the normal distribution. In addition, the standard deviation was big in comparison with the mean. These statistics reinforce the idea that there are some individuals with extreme values. Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some high values (right tail), probably clients with 4 or 5 calls.
* **Average Monthly GB Download:** the average number of downloaded GB per month is approximately 6.7 GB. The 50% of the subscribers download 5GB per month or less. Moreover, the majority of the subscribers never downloaded films or series from the service, as the mode is equal to 0. Probably the streaming option is the favourite. On the other hand, data is really dispersed: standard deviation is equal to 7.45 GB, meaning that there are some atypically extreme values, most probably customers that have downloaded a lot of GB per month versus customers with no downloads at all. The range is very wide: there are clients which have downloaded 0 GB per month, other/s up to 43 GB per month. The positive kurtosis shows that values are more dispersed from the mean and with stronger tails than the normal distribution. In addition, the standard deviation was big in comparison with the mean. For all the above reasons, it is clear there are some individuals with extreme values. Finally, positive symmetry also confirms that individuals are concentrated in lower values and there are some high values (right tail).
* **Extra Data Charges:** the average extra data charges is almost 3.4 USD. Moreover, the majority of the subscribers have not been charged extra as the mode is equal to 0 USD. Same for the median. Most probably, this variable is positively correlated with the above (average monthly GB download): the more the GB downloaded, the more the extra data charge. Check this. On the other hand, data is really dispersed: standard deviation is equal to 12.59 USD (3.7 times the mean) meaning that there are some extreme values, most probably customers that had been charged extra due to big downloads of GB versus customers with no charges at all. This is confirmed by the range value of 99 USD: there are clients which had been charged up to 99 USD dollars and clients with no charges. The extremely positive kurtosis shows that values are much more dispersed from the mean and with much heavier tails than the normal distribution. In addition, the standard deviation was much bigger in comparison with the mean. Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some high values for the extra data charge (right tail).
* **Age:** the average age of the subscribers is equal to 47.53 years old. The median is very similar: 47 years old. In addition, there is a majority of subscribers who are 29 years old (mode). However, data dispersion is notorious: the standard deviation is approximately 17 years old. This idea is reinforced when we observe the range of dates: we have subscribers who are 19 and up to 85 years old. It is clear that Amazon Prime Video offers a variety of contents suitable for all ages. Moreover, the negative kurtosis suggests that values are concentrated around the mean (lighter tails than the normal distribution). However, the standard deviation was big in comparison with the mean. These statistics reinforce the idea that there are some individuals with very atypical values (outliers). Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some atypically high values (right tail).
* **Number of customers in Group**: the average number of customers in Group is less than 1. Moreover, the majority of the subscribers are enrolled in an Individual Plan as the mode is equal to 0 USD. Same for the median. On the other hand, data dispersion is a fact: standard deviation is equal to 1.7 (2 times the mean) meaning that there are some extreme values, most probably customers that are enrolled in a numerous Family Plan coexist with customers in Individual plans. This is confirmed by the range value of 6. The positive kurtosis shows that values are more dispersed from the mean and with heavier tails than the normal distribution. Finally, positive symmetry suggests that individuals are concentrated in lower values (0 – 2) and there are some high values (right tail) (5 – 6).
* **Monthly Charge (USD):** the average monthly charge is equal to 30.90 USD/month. The median is very similar: 31 USD/month. In addition, the majority of subscribers pay a charge of 10 USD/month. However, data dispersion is notorious: the standard deviation is approximately 16.40 USD. This idea is reinforced when we observe the range of monthly charges (78): we have subscribers who pay 0 USD/month in charges and others who pay up to 78 USD/month. Moreover, the negative kurtosis suggests that values are concentrated around the mean (lighter tails than the normal distribution). However, the standard deviation was big in comparison with the mean. These statistics reinforce the idea that there are some individuals with very atypical values (outliers). Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some atypically high values (right tail).
* **Total charges (USD):** the average total charges is equal to 1084.78 USD. The median is much lower: 647 USD, meaning that the 50% of subscribers paid a total of 647 USD or less. In addition, the majority of subscribers pay a charge of 10 USD/month. However, data dispersion is notorious: the standard deviation is approximately 16.40 USD. This idea is reinforced when we observe the range of monthly charges (78): we have subscribers who pay 0 USD/month in charges and others who pay up to 78 USD/month. Moreover, the negative kurtosis suggests that values are concentrated around the mean (lighter tails than the normal distribution). However, the standard deviation was big in comparison with the mean. These statistics reinforce the idea that there are some individuals with very atypical values (outliers). Finally, positive symmetry suggests that individuals are concentrated in lower values and there are some atypically high values (right tail).