Project Outline:

1. Data Cleaning and exploration
2. Data visualisation
3. Data Interpretation (Order Processing Analysis)

3.1. Investigate how orders are currently being processed, so the times it takes between different appointments.

-> Are there any noticeable bottlenecks?

-> Can you identify any trends or patterns that change over time?

3.2. Additional Insights: Beyond the provided guidelines, we encourage you to dive deeper and bring to light any additional insights you think could be beneficial to the business.

To Dos:

1. Check for missing values
2. Check for null values
3. Duplicates
4. Data Types ✓
5. Other inconsistencies (orders\_df[‘postal\_code’] has nan values, replace by “ “
6. Check how many rows we have in each dataset. ✓
7. definition of each variable

Other points:

Keep the business in mind.

As mentioned in the interview:

1. Partner Networking: When a person buys a new house, we offer him services. -> Higher conversion rate.

* How do we generate revenue?
* What are our (strongest) operational divisions?

Product / operational data

Data

Suppose you have a dataset of customer orders. You might want to:

* Check that each order\_id is unique to ensure that no orders are recorded more than once.
* Identify if any customer has placed identical orders multiple times, which might indicate a duplicate entry.
* Verify the unique categories in a product\_type column to ensure all expected product types are present.

**Summary**

* **Unique Values**: Help understand the diversity and ensure the expected variety in the data.
* **Duplicated Values**: Help identify and rectify redundant or erroneous entries.

**Interpretation of the Boxplots**

1. **On-site Appointment:**
   * The median (middle line) is around 40 days.
   * The interquartile range (IQR) extends from approximately 25 to 55 days.
   * There are several outliers beyond 80 days, with a few extending to around 140 days.
2. **Project Call:**
   * The median is around 50 days.
   * The IQR extends from approximately 30 to 70 days.
   * There are numerous outliers beyond 80 days, reaching up to 140 days.
3. **Final Call:**
   * The median is around 55 days.
   * The IQR extends from approximately 35 to 75 days.
   * Several outliers are beyond 80 days, with some as high as 160 days.

**Derivations from the Boxplots**

* **Central Tendency:**
  + The median time to the final call is slightly higher than for the other two appointment types.
* **Variability:**
  + The spread of the data (IQR) is similar for all appointment types, indicating consistent variability across the three types.
* **Outliers:**
  + There are significant outliers in all appointment types, suggesting that some orders take much longer than the typical range.

**Handling Outliers**

1. **Identify the Cause:**
   * Investigate why these outliers exist. Are they due to specific conditions or errors in data recording?
2. **Transformation:**
   * Apply transformations (e.g., log transformation) to reduce the impact of outliers.
3. **Filtering:**
   * Remove outliers based on a chosen threshold (e.g., 1.5 times the IQR above the third quartile and below the first quartile).
4. **Imputation:**
   * Replace outliers with a statistical measure like the median or mean of the data.
5. **Robust Statistical Methods:**
   * Use statistical methods that are robust to outliers, such as the median instead of the mean for central tendency.

* Why do we use the median here?

Interpreting the boxplot:

I am using a Boxplot to understand the relationship between a categorical variable (on-site-appointment) and a continuous variable (number of days from order date). i want to interpret this. the visual shows outliers. I want to explain how long the average days are to have the on-site-appointment. do i use the median or mean=

* When interpreting boxplots and explaining the central tendency of your data, it is typically more appropriate to use the median rather than the mean. This is because the median is less affected by outliers and provides a better representation of the typical value in skewed distributions.

Summary Statistics Interpretation

For on-site appointments:

Count: 17,265 (number of observations)

Mean: 27.27 days (average number of days from the order date)

Standard Deviation: 20.11 days (variability in the number of days)

Min: 0 days (minimum number of days)

25th Percentile (Q1): 15 days (25% of the appointments are scheduled within 15 days)

Median (Q2, 50th Percentile): 21 days (middle value, 50% of the appointments are scheduled within 21 days)

75th Percentile (Q3): 33 days (75% of the appointments are scheduled within 33 days)

Explanation for On-site Appointments

To explain how long it typically takes to schedule an on-site appointment:

Median (Central Tendency): The median number of days from the order date to the on-site appointment is 21 days. This is a robust measure of central tendency, not affected by outliers.

Interquartile Range (IQR): The IQR is from 15 to 33 days. This range indicates that 50% of the appointments are scheduled within 15 to 33 days from the order date.

Mean (Average): The average number of days is 27.27. However, because the mean can be affected by outliers, it might not be as representative of the typical scheduling time as the median.

Outliers: The standard deviation is 20.11 days, indicating there is some variability in the scheduling times. Outliers are present, as seen in the boxplot, and they can significantly extend the scheduling time beyond the typical range.

Handling Outliers

Given the presence of outliers, here's how you might handle and explain them:

Explaining Outliers:

Appointments taking more than 60 days from the order date can be considered outliers. These cases should be investigated to understand the reasons for such delays.

Reporting:

When reporting the typical scheduling time, focus on the median and IQR.

Mention that while the median time is 21 days and most appointments (50%) are scheduled within 15 to 33 days, there are some instances where it takes significantly longer, as indicated by the outliers.

Summary for Reporting

For on-site appointments, the median number of days from the order date is 21 days, with an interquartile range of 15 to 33 days. This means that half of the appointments are scheduled within this range. Although the average time is 27.27 days, this figure is influenced by outliers, with some appointments taking over 60 days to schedule. These outliers should be further investigated to understand and potentially mitigate the causes of such delays.