**Project Description:**

In response to the challenge of understanding loan repayment behaviors among urban applicants, our team conducted a comprehensive **Exploratory Data Analysis (EDA)** using real-world loan application data. The primary objective of this project was to uncover meaningful patterns that differentiate loan defaulters from those who successfully repay, thereby supporting more informed and data-driven lending decisions.

This analysis explored key factors such as **applicant income, credit amount, occupation, education level, years of employment**, and other demographic and financial variables. The dataset was **segmented based on repayment status**—'Default' and 'Repaid'—to allow focused investigation of both groups.

Using Excel-based techniques like **pivot tables, correlation analysis, and visual dashboards**, we identified high-impact indicators associated with loan default and repayment. Strong correlations were observed between loan amounts, annuities, and goods prices across both segments, while occupation and education levels provided deeper insights into applicant profiles.

The ultimate goal of this initiative is to aid in the development of a more refined and fair loan evaluation framework—**one that minimizes the risk of default without excluding creditworthy applicants**. Through this data-driven approach, the company aims to enhance financial sustainability while better serving its urban customer base.

**Tech Stack used:**

The data cleaning, visualization, and analysis tasks were carried out using the Excel 2021 version. Additionally, MS Word was used for generating reports.

**Task A – Identify Missing Data and Deal with It Appropriately**

**Introduction:**

As part of our data analysis initiative to better understand loan default risks among urban customers, we began by examining and addressing missing data in the loan application dataset. Handling missing data is essential to maintain the accuracy and integrity of any analysis. This section describes the steps we followed to identify and clean missing data using Excel functions and Power Query, supported by visualizations to illustrate the extent of missing values.

**Task Overview:**

The main goal was to detect missing values in the dataset and handle them effectively. We used Excel functions such as COUNTBLANK and IF to identify missing entries and applied suitable imputation methods. Additionally, we performed several ETL (Extract, Transform, Load) operations using Power Query to clean and reshape the dataset before deeper analysis.

**Methodology:**

**1. Data Cleaning with Power Query:**

We used Power Query for several initial data preparation tasks:

* Removed unnecessary columns that were not relevant to our analysis.
* Renamed columns for better readability.
* Filtered rows to exclude records with incomplete or irrelevant data.
* Created two new columns:
  + AGE from DAYS\_BIRTH (converted from days to years).
  + YEARS\_EMPLOYED from DAYS\_EMPLOYED (converted from days to years).
* Ensured all date and numeric columns were properly formatted.

These transformations helped make the dataset cleaner and more meaningful before proceeding with further analysis.

**2. Missing Data Identification:**

* We used the COUNTBLANK function to calculate the number of blank cells in each column.
* To get the percentage of missing values, we divided the count of blanks by the total number of rows and multiplied by 100.
* The IF function was used to classify whether a column had missing values or not.

This allowed us to identify which columns had high levels of missing data.

**3. Handling Missing Values:**

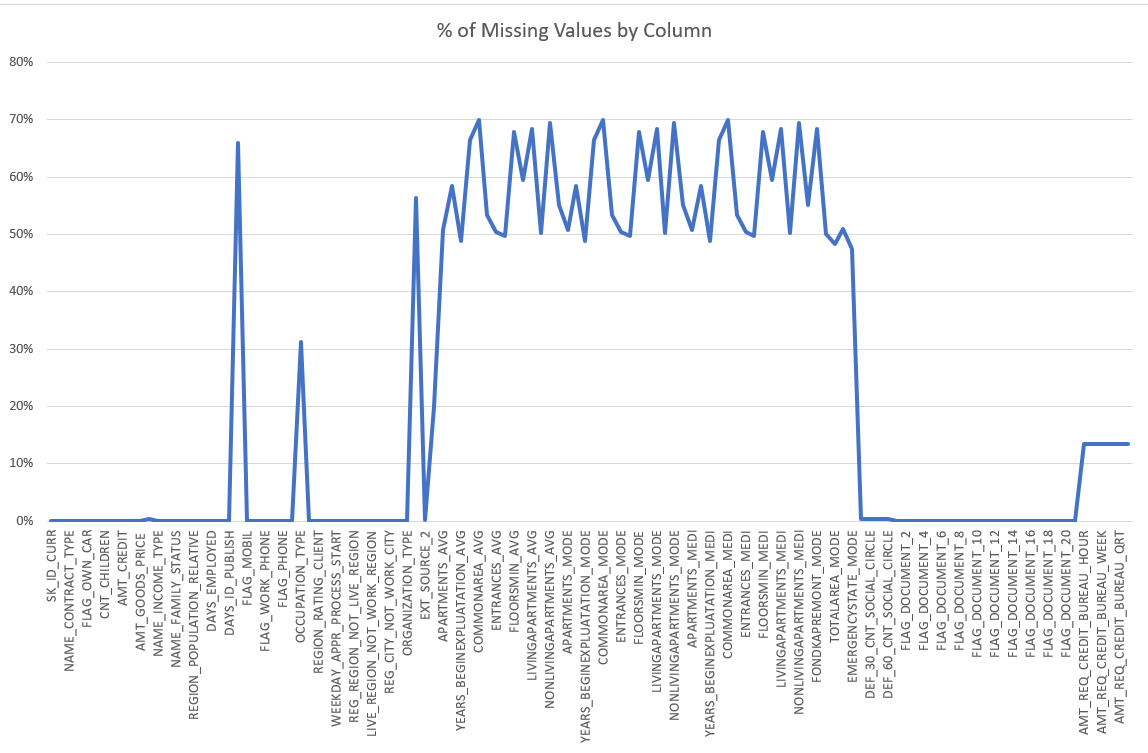
* Columns with a very high percentage of missing values and irrelevant columns were dropped to avoid skewed analysis.
* For other columns with only a small percentage of missing values, we used the following methods:
  + AVERAGE function to fill missing values for numerical variables when the data was normally distributed.
  + MEDIAN function when the data had outliers or was skewed.

These imputation strategies helped us maintain the structure and meaning of the data without introducing bias.

**Results and Visualizations:**

To summarize our findings:

* A line chart was created showing the percentage of missing values across the key variables.
* This visual representation gave us a clear understanding of where the most data loss occurred and guided our cleaning decisions.



**Task B: Identify Outliers in the Dataset**

**Introduction:**  
Understanding how loan amounts and financial obligations vary among applicants is crucial for identifying unusual patterns that may impact loan repayment. In this task, we focused on visualizing outliers in key financial variables to get a clearer picture of the data spread and identify any extreme values.

**Task Overview:**  
The objective of this task was to identify and visualize outliers in selected amount-related columns. We did not remove any outliers from the dataset. Instead, we used Excel charts to explore the distribution of values and highlight significant deviations.

**Methodology:**

1. Columns Used for Outlier Visualization:  
   We selected the following financial variables for outlier analysis:

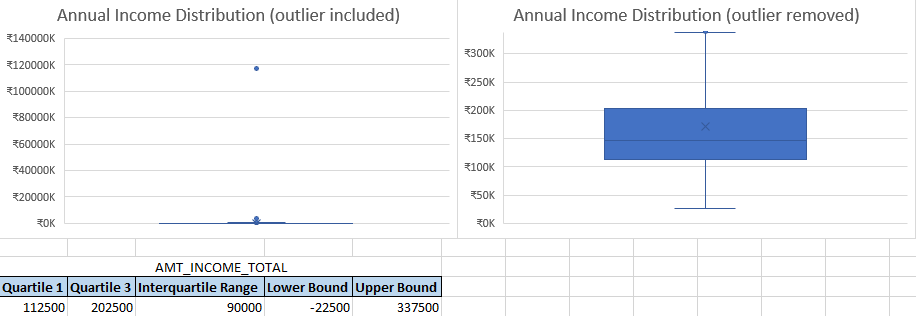
* AMT\_INCOME\_TOTAL
* AMT\_CREDIT
* AMT\_ANNUITY
* AMT\_GOODS\_PRICE

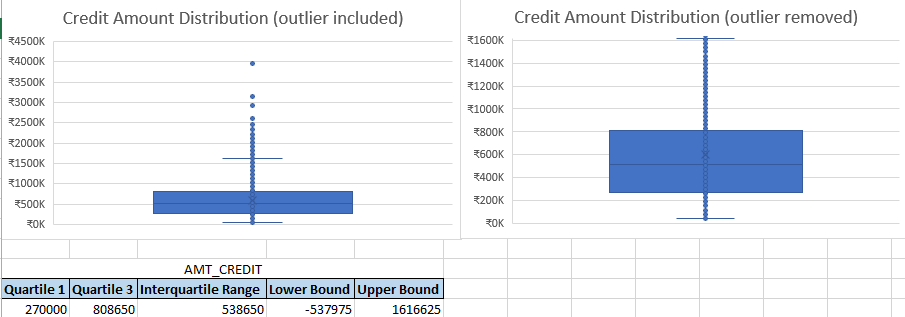
1. Visualization Technique:

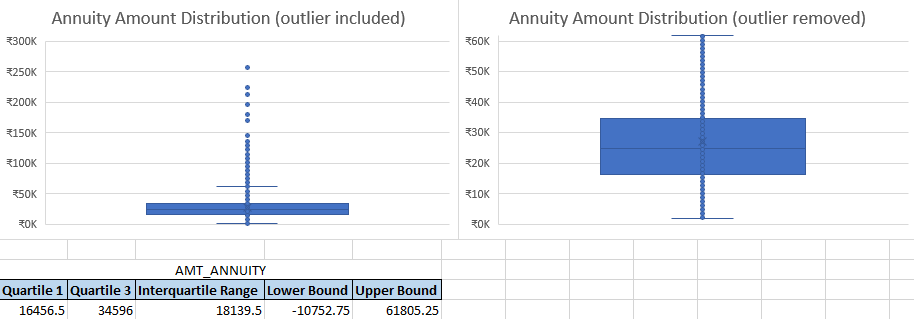
* We used box plots in Excel to visualize the distribution of each amount column.
* These plots clearly displayed the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values.
* Any data points that appeared outside the whiskers (typically 1.5 times the IQR) were considered potential outliers.

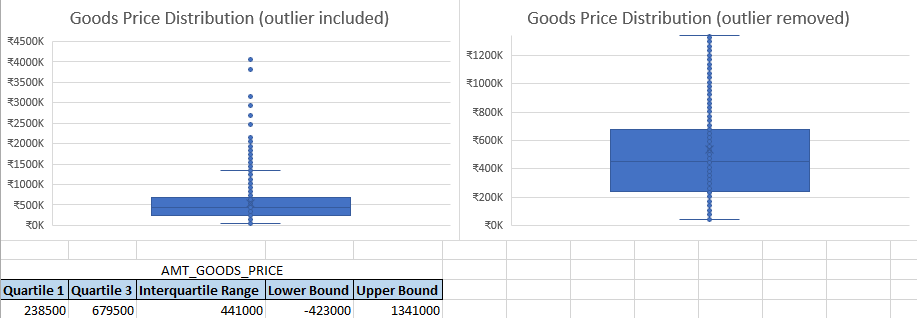
1. ETL and Power Query Usage:

* We used Power Query to prepare the dataset before visualization.
* Only relevant amount columns were retained for this task.
* We ensured that all selected fields were properly formatted as numerical values for accurate plotting.

**Results and Visualizations:**  
Box plots were successfully created for each of the selected amount columns. These visualizations helped us identify several outlier values, particularly in AMT\_INCOME\_TOTAL and AMT\_CREDIT, where a small number of applicants had unusually high figures. These extreme values were not removed but noted for their potential impact ****on further analysis**.**

****

****

****

**Task C: Analyze Data Imbalance**

**Introduction:**  
As part of our analysis on urban loan defaults, we examined the distribution of loan repayment outcomes. Understanding how balanced the dataset is between customers who repaid and those who defaulted is important to ensure fair representation during our data exploration.

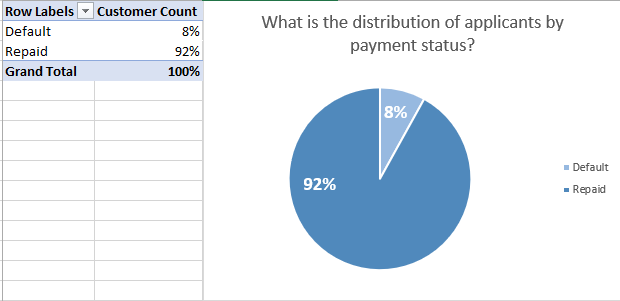
**Task Overview:**  
The objective was to analyze the proportion of loan repayment statuses to identify any imbalance in the dataset. We used Excel Pivot Tables and a Pie Chart to visualize the distribution.

**Methodology:**

1. Data Aggregation:  
   We used a Pivot Table to calculate the total number of customers under each repayment status: Repaid and Defaulted.
2. Visualization:  
   A Pie Chart was created based on the pivot summary to easily see the proportions of each class.
3. Findings:

* 92% of the customers had repaid their loans.
* Only 8% had defaulted.

**Results and Visualisations:**  
The Pie Chart visually confirmed that the dataset is imbalanced, with a large majority of the records belonging to the Repaid category. This indicates that default cases are underrepresented in the data.



**Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis**

**Task D.1: Univariate Analysis**

**Introduction:**  
Univariate analysis is the examination of a single variable in isolation. It helps identify basic distributions, frequencies, and proportions within the dataset. In this task, we focused on understanding the distribution of customer occupations.

**Task Overview:**  
The objective was to analyze the distribution of the occupation variable to identify the most common customer segments and any data quality issues.

**Methodology:**

1. **Data Aggregation**:

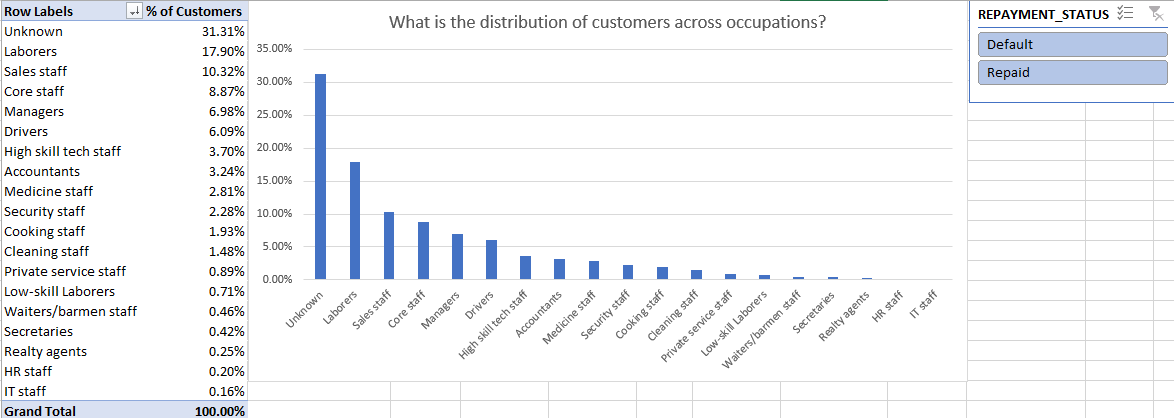
A Pivot Table was created in Excel to summarize the percentage of customers by occupation type.

1. **Visualization**:

* A Column Chart was generated to visually represent the distribution of customers across occupations.

**Results and Visualizations**:

* 31.31% of customers have an occupation listed as "Unknown".
* The next most frequent occupations are Laborers (17.90%) and Sales staff (10.32%).
* Notably, 25% of defaulters belong to the "Unknown" occupation group, indicating a high risk associated with missing occupational data.
* The column chart clearly visualizes these proportions.



**Insight**:  
The large percentage of customers with unknown occupation highlights a significant data collection issue. Incomplete occupation data limits the ability to accurately assess risk and identify default patterns, underscoring the need for improved data quality processes.

**Task D.2: Segmented Univariate Analysis**

**Introduction:**  
Segmented univariate analysis involves examining the distribution of variables within different segments to reveal deeper insights about patterns influencing loan repayment.

**Sheet:** Age vs Repayment

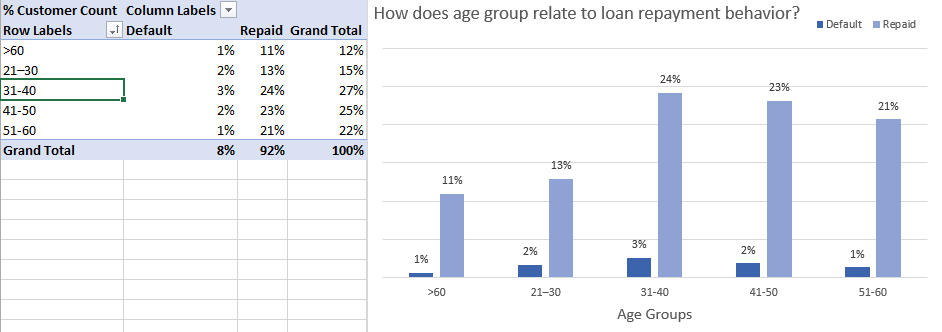
**Task Overview:**  
Analyze the distribution of customer age groups segmented by repayment status to understand which age brackets have higher default rates.

**Methodology:**

* Used a pivot table to calculate percentage distribution of customers by age group and repayment status.
* Visualized results with a clustered column chart for clear comparison.

**Results and Visualizations:**

* The age groups 31-40 (27%) and 41-50 (25%) have the largest share of customers.
* Default rate is highest in the 31-40 age group (3%).
* Youngest (>21-30) and oldest (>60) groups have lower default proportions.



**Insight:**  
Middle-aged customers tend to have a higher share of defaults. This suggests age influences repayment behavior and can be an important factor in loan approval decisions.

**Sheet:** Credit Amount by Region Rating

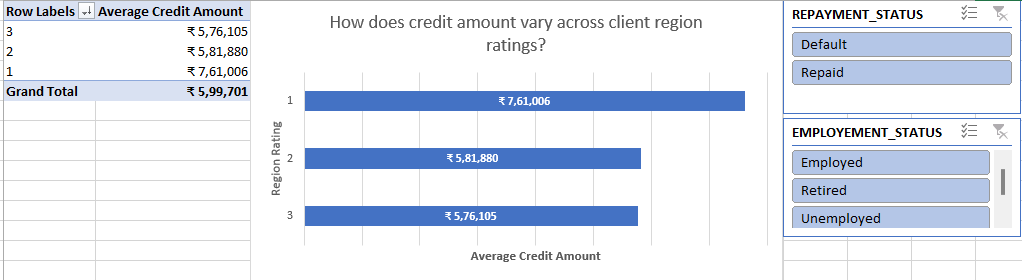
**Task Overview:**  
Examine average credit amount segmented by region rating and repayment & employment status to identify patterns in credit amount approval for different customer segments.

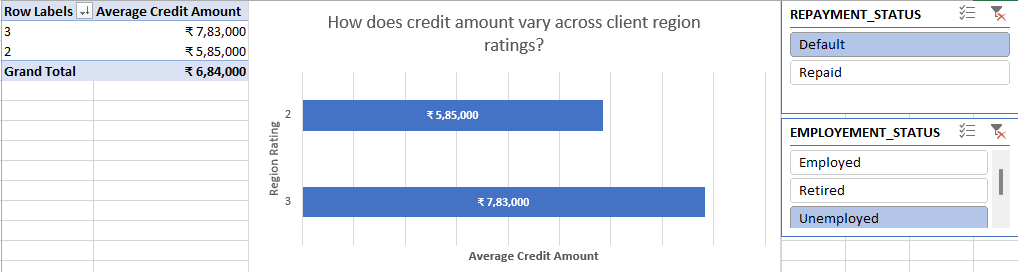
**Methodology:**

* Used a pivot table with slicers applied to focus on defaulted and unemployed customers.
* Calculated average credit amount by region rating.
* Visualized using bar charts.

**Results and Visualizations:**

* For all customers, credit amount is highest in region rating 1 (~₹7.6 lakh).
* Among unemployed defaulters, average credit amount in region 3 rises to ₹7.83 lakh, higher than other ratings.





**Insight:**  
Certain region ratings, particularly among unemployed defaulters, are associated with higher credit amounts, indicating riskier lending patterns in these segments.

**Sheet:** Credit Amount vs Education

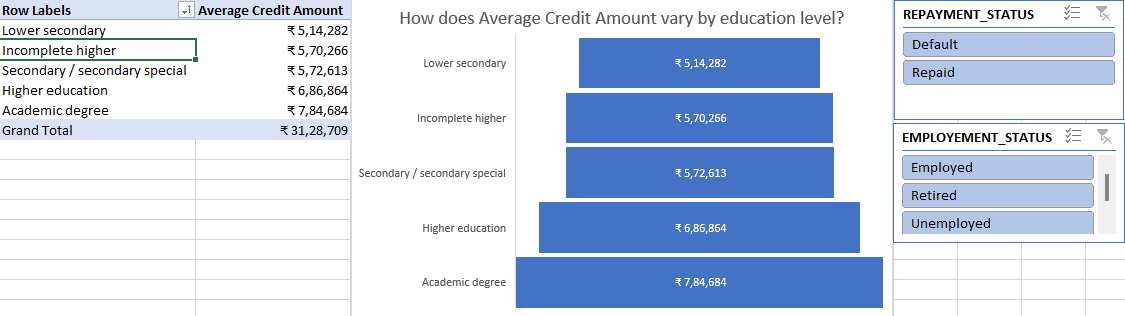
**Task Overview:**  
Assess how average credit amount varies by education level segmented from the whole dataset.

**Methodology:**

* Pivot table to calculate average credit amount by education category.
* Funnel chart to visualize the distribution in decreasing order.

**Results and Visualizations:**

* Average credit amount increases with education level, from ₹5.14 lakh (Lower secondary) to ₹7.84 lakh (Academic degree).
* Customers with higher education tend to receive larger loan amounts.

****

**Insight:**  
Education appears to positively influence loan amount granted, potentially reflecting income or creditworthiness associated with educational attainment.

**Task D.3: Bivariate Analysis**

**Sheet:** Income by Gender and Repayment

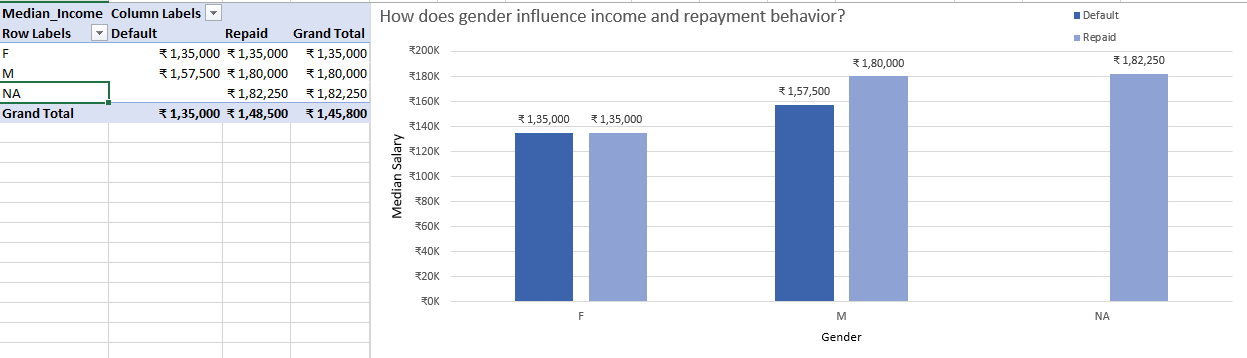
**Task Overview:**  
Explore the relationship between income, gender, and repayment status to identify patterns influencing loan repayment.

**Methodology:**

* Pivot table used to compute median income by gender and repayment status.
* Clustered column chart to compare income across groups.

**Results and Visualizations:**

* Median income for males who repaid (₹1,80,000) is higher than males who defaulted (₹1,57,500).
* Female income remains constant at ₹1,35,000 regardless of repayment.
* Customers with unknown gender have higher median income when repaid.



**Insight:**  
Higher income correlates with better repayment among males. Gender combined with income can provide insights for credit risk profiling.

**Sheet:** Income vs Credit

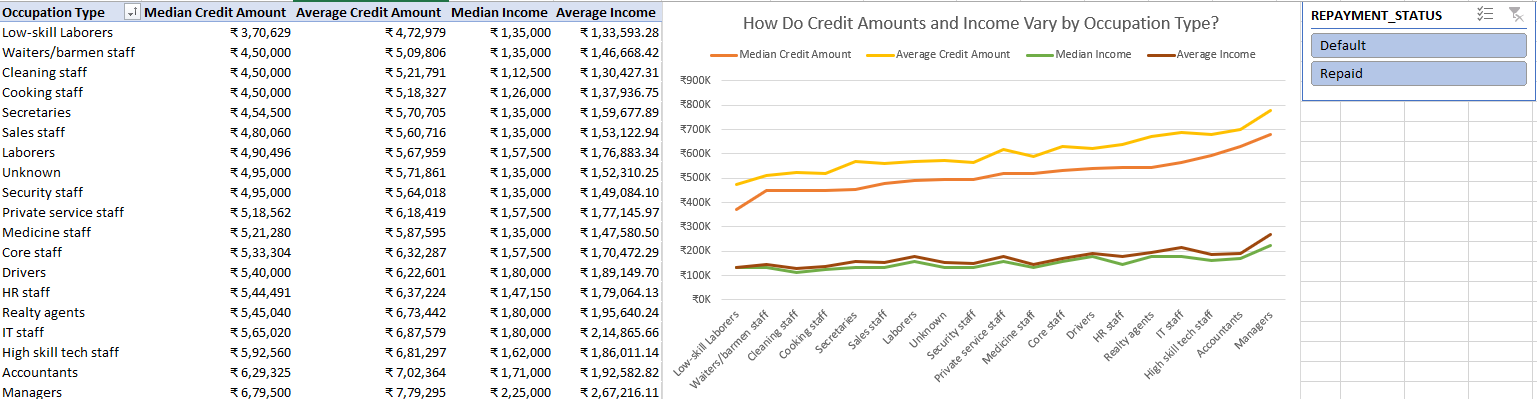
**Task Overview:**  
Analyze the relationship between occupation type, credit amount, and income to identify occupation groups with higher credit amounts and income.

**Methodology:**

* Pivot table with occupation as rows, and median/average credit and income values as columns.
* Line chart to visualize trends across occupations.

**Results and Visualizations:**

* Occupations like Managers, Accountants, and IT staff have higher median incomes and credit amounts.
* Lower-skilled laborers and unknown occupations show lower income and credit values.



**Insight:**  
Occupation type strongly influences credit amount and income, which are key indicators for loan repayment capability.

**Task E: Identify Top Correlations for Different Scenarios**

**Introduction:**  
Correlation analysis helps uncover relationships between numerical variables, providing insights into which customer or loan features are strongly associated with each other. For this task, we focused on identifying the top correlations within two distinct customer segments—those who defaulted on their loans and those who successfully repaid—to understand which attributes are most interconnected in each scenario.

**Task Overview:**  
The objective of this task was to calculate and compare correlation matrices across two repayment outcome segments: clients with payment difficulties and clients with successful repayment histories. This comparison highlights which features move together and may serve as key indicators of risk or reliability.

**Methodology:**

1. **Data Segmentation:**
   * The dataset was split into two segments using filters in Excel:
     + Clients with payment difficulties (defaults)
     + Clients who repaid loans successfully
2. **Correlation Calculation:**

* Using Excel’s CORREL function, we computed pairwise correlation coefficients for relevant numerical variables:
  + - AGE
    - CNT\_CHILDREN
    - AMT\_INCOME\_TOTAL
    - AMT\_CREDIT
    - AMT\_ANNUITY
    - AMT\_GOODS\_PRICE
    - YEARS\_EMPLOYED

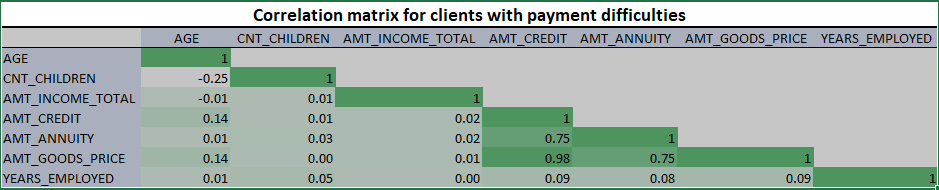
1. **Visualization:**

* Correlation matrices were created for each customer segment.
* High correlation pairs were highlighted for easier comparison.

**Results and Insights:**

**For Clients with Payment Difficulties:**

* The strongest correlation observed was between AMT\_CREDIT and AMT\_GOODS\_PRICE (r = **0.98**), followed by AMT\_CREDIT and AMT\_ANNUITY (r = **0.75**).
* AGE showed weak or no correlation with income or credit-related attributes.
* CNT\_CHILDREN had a weak negative correlation with AGE (r = **-0.25**), suggesting younger customers may have more dependents in this segment.



**For Clients with Successful Repayment:**

* Again, AMT\_CREDIT and AMT\_GOODS\_PRICE had the highest correlation (r = **0.99**).
* Moderate-to-strong positive correlations were observed between:
  + AMT\_INCOME\_TOTAL and AMT\_ANNUITY (r = **0.47**)
  + AMT\_INCOME\_TOTAL and AMT\_GOODS\_PRICE (r = **0.41**)
* CNT\_CHILDREN had a stronger negative correlation with AGE (r = **-0.33**) compared to the default segment.

