Data Analyst Portfolio

Trainity Data Analytics Internship

Deepthy A

Professional Background

Hello, I'm Deepthy A - a curious mind with a strong passion for transforming raw data into meaningful insights. With a natural flair for analytical thinking and a growing foundation in business decision-making, I embarked on my data journey through the Trainity Data Analytics Internship.

Throughout this experience, I've worked hands-on with tools like Excel, SQL, Power BI, and data visualization techniques, diving into projects that mirrored real-world challenges across industries like finance, marketing, operations, and customer service.

From scrubbing messy datasets to uncovering hidden patterns, I found my strength in:

- **Data cleaning that improves reliability,**
- **Exploratory analysis that tells stories,**
- **Interactive dashboards that inform decisions, and**
- **Actionable insights that move the needle for businesses.**

Each project deepened my understanding of how data can drive smarter choices — and confirmed my excitement for this field. I'm now ready to bring this passion and skillset into a professional setting, where I can contribute to meaningful, data-driven change.

Table of Contents

Using Data Analytics in Real Life: Shopping Decision-Making Example

Context:

A relatable, everyday scenario - shopping for clothes during a seasonal sale. Whether online or in-store, we often make decisions influenced by personal preferences, budget, past experience, and available options. This everyday process is actually driven by key data analytics principles.

Description:

Shopping might seem like a simple task, but it involves multiple levels of decision-making that mirror the data analytics process. From identifying needs to analyzing trends and comparing options, every step reflects how we use available data to make informed and smart decisions.

Key Elements:

- Identifying the shopping goal (new outfit for a festival or event)
- Budget planning and constraints
- Researching options (brands, prices, offers)
- Analyzing what combinations work well (based on color, fit, past purchases)
- Taking recommendations or reviews into account
- Final purchase and reflection

What’s Involved?

1**. Plan**:

It all starts with a simple goal-finding the right outfit for something like a family function or get-together. At this point, there's usually an idea in mind about the overall vibe, like something semi-formal or casual. Once that’s clear, the focus shifts to figuring out exactly what’s needed-maybe a new shirt, a nice pair of jeans, or a couple of accessories to pull the look together.

2. **Prepare:**

Before diving into shopping, the budget is set. It’s also common to quickly glance through the wardrobe to avoid buying something that’s already there. Along the way, a bit of research helps-like checking for ongoing discounts, looking at store timings, or browsing delivery options if shopping online.

3. **Process:**

Now comes the actual searching part. Whether it’s scrolling through websites or walking through stores, there’s a lot to compare-colours, styles, materials, brands, prices. Some options are ruled out right away, like if the size isn’t available or the design doesn’t match personal taste. It’s basically about filtering the noise to get to the good stuff.

4. **Analyze:**

Once a few options are in hand, the next step is figuring out what works best. That could mean matching a new shirt with jeans already owned or checking if the color combo is in trend. Reviews, style blogs, or a quick opinion from a friend help make sure the final choice looks and feels right.

5. **Share:**

Getting a second opinion is always part of the mix-sending photos to friends or asking a store assistant for suggestions. These small bits of feedback go a long way in building confidence in the choice and spotting things that might’ve been missed.

6**. Act:**

With everything lined up, it’s time to go ahead and make the purchase. If everything clicks-budget, style, comfort-it feels like a win. And usually, there’s a moment after buying where it’s all mentally reviewed: was it worth it, is it going to be worn often, or was it just an impulse? These reflections naturally shape how similar decisions are made in the future.

Real-Life Impact:

- Helps avoid unnecessary spending while still getting what’s needed.
- Makes sure new items actually go well with what’s already in the closet.
- Leads to more satisfying purchases by mixing logic, style, and budget.
- Builds up shopping confidence over time through small, thoughtful choices.

Instagram User Analytics

SQL Project

Project Description

This analysis investigates how individuals utilize Instagram by executing SQL queries against an Instagram clone database. The objective is to reveal insights that assist the marketing team in rewarding loyal users, re engaging inactive users, selecting winning content strategies, suggesting top hashtags, determining the best day for ad campaigns, and detecting suspicious "bot like" behavior.

We will use SQL and MySQL Workbench to query user behavior data to solve real business problems:

- **Who are our longest-standing users?** (Loyalty rewards)
- **Which accounts have never posted?** (Re-engagement campaigns)
- **Whose photo became most popular?** (Contest winner)
- **Which hashtags drive the most traction?** (Marketing recommendations)
- **When do most people sign up?** (Optimal ad timing)
- **How active is the average user?** (Engagement metrics)
- **Are there bot-like accounts lurking?** (Fake account detection)

Approach

We tackled this in four clear phases:

1. **Database Setup:**
    - Loaded the provided SQL script into MySQL Workbench.
    - Executed it to build tables and populate them with sample data.

2. **Query Development:**
    - For each business question, wrote a focused SQL statement. We combined JOINs, LEFT JOINs, GROUP BY, and HAVING clauses to filter and aggregate the data.
    - Used subqueries where needed (for example, to compare a user's like count against the total number of photos).
3. **Result Validation & Capture:**
    - Ran each query and double-checked counts with simple tests (e.g., SELECT COUNT(*) FROM photos;).
    - Captured screenshots of the SQL editor showing both the query and its result grid. These live examples ensure transparency and reproducibility.
4. **Insight Synthesis:**
    - Interpreted the raw numbers to pull out key takeaways—like which day of the week is peak registration time.
    - Framed these findings in a way that ties back to Instagram's goals: growing engagement, rewarding users, and optimizing ad spend.

Tech Stack Used

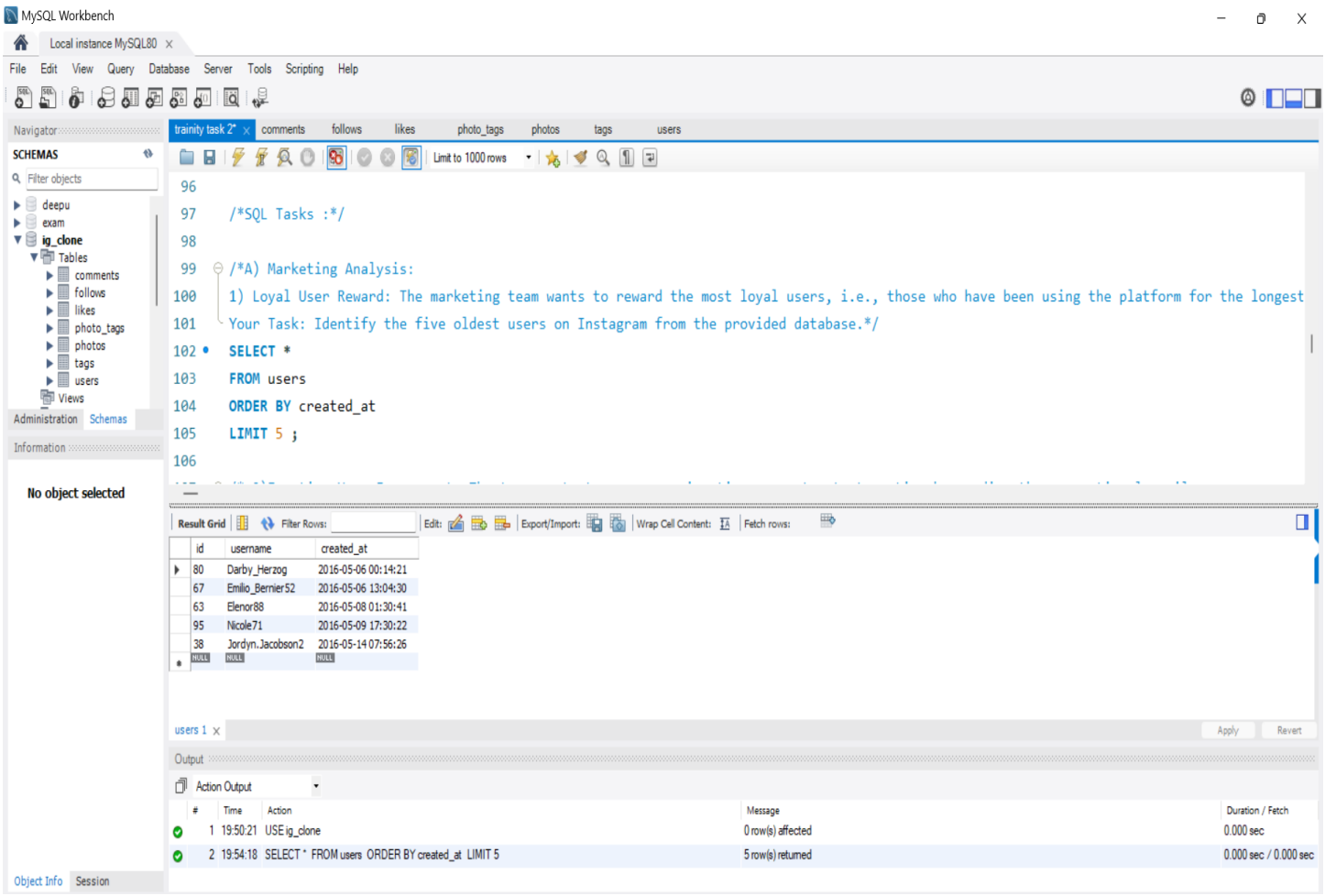
Tool / Library	Version	Purpose
MySQL Server	8.x	Hosts the relational database engine.
MySQL Workbench	8.x	Schema design, SQL editing, output visualization.
Microsoft Word	365	Report drafting, embedding screenshots, annotations.

Why MySQL Workbench?

Its intuitive GUI for writing, formatting, and running queries made it effortless to iterate on SQL and immediately see the results. The visual explain plans also helped optimize longer queries.

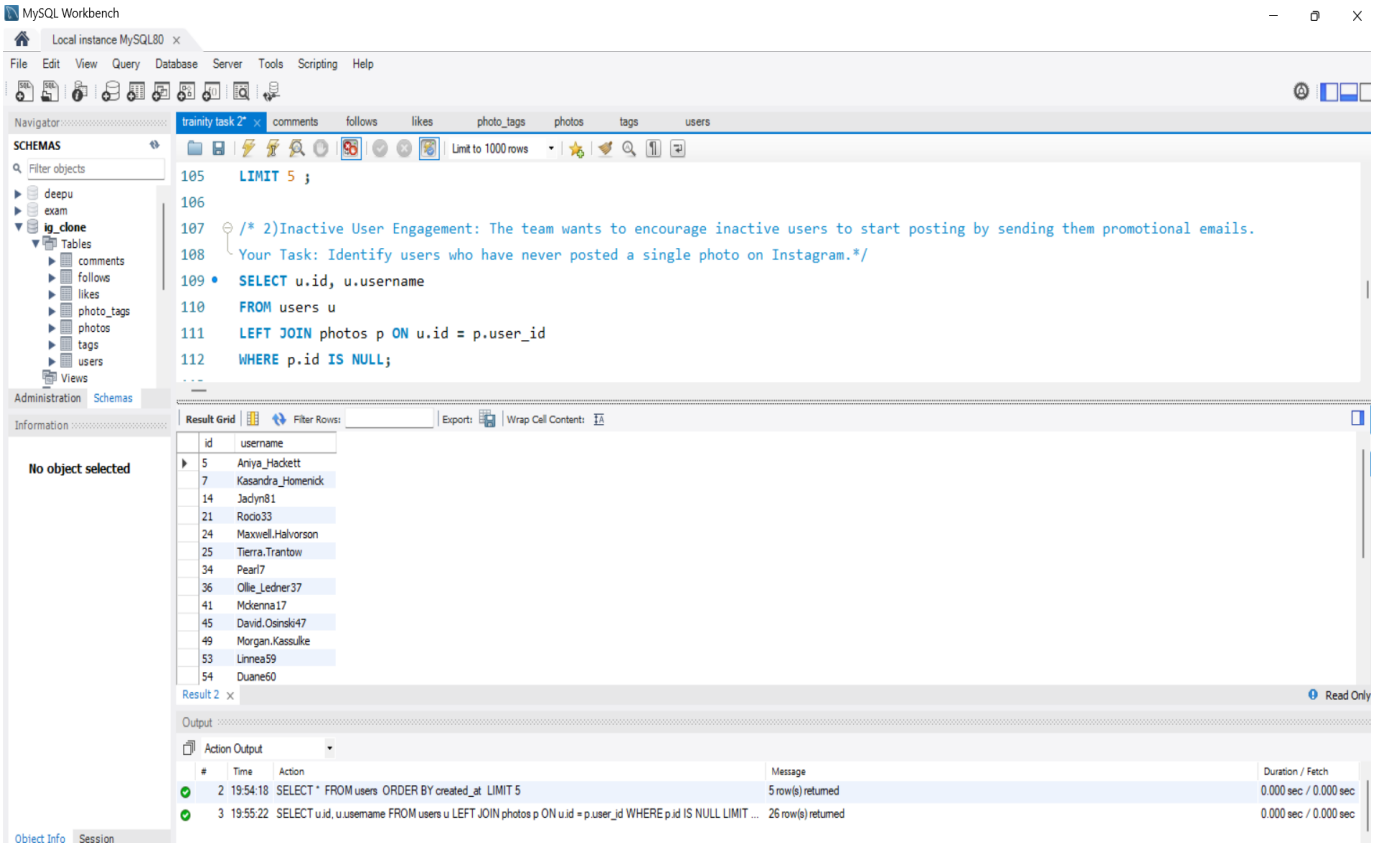
SQL Queries & Outputs

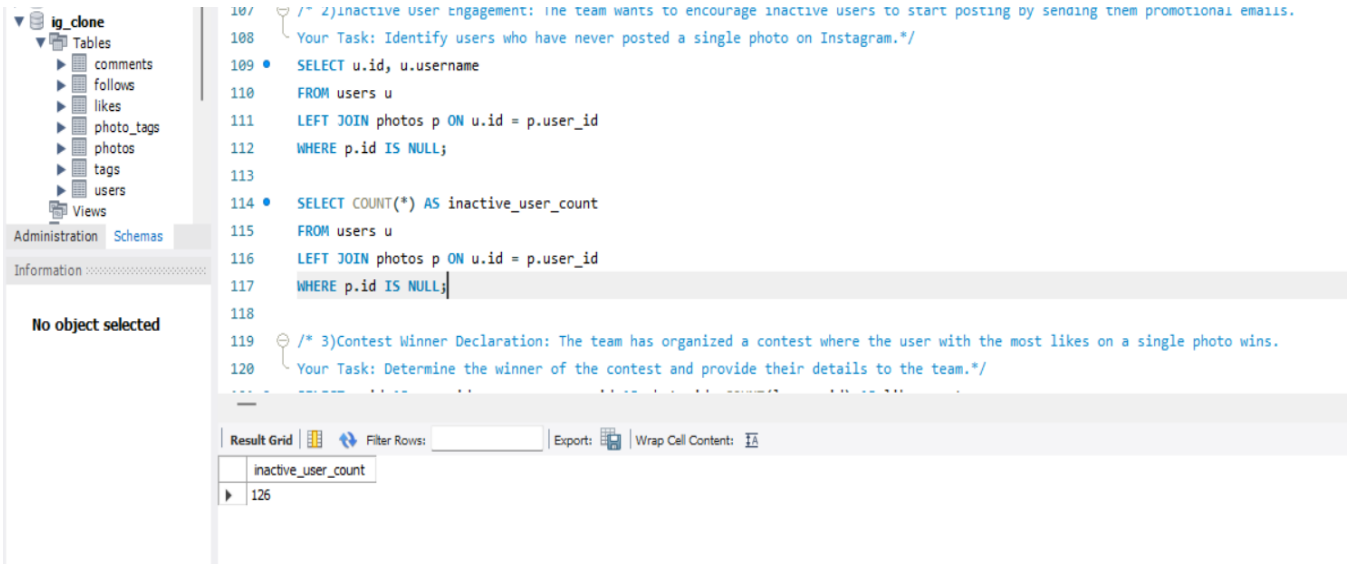
Task 1: Loyal User Reward



Key takeaway: These five “veteran” users have been part of the community the longest—ideal candidates for special loyalty perks or early-adopter recognition.

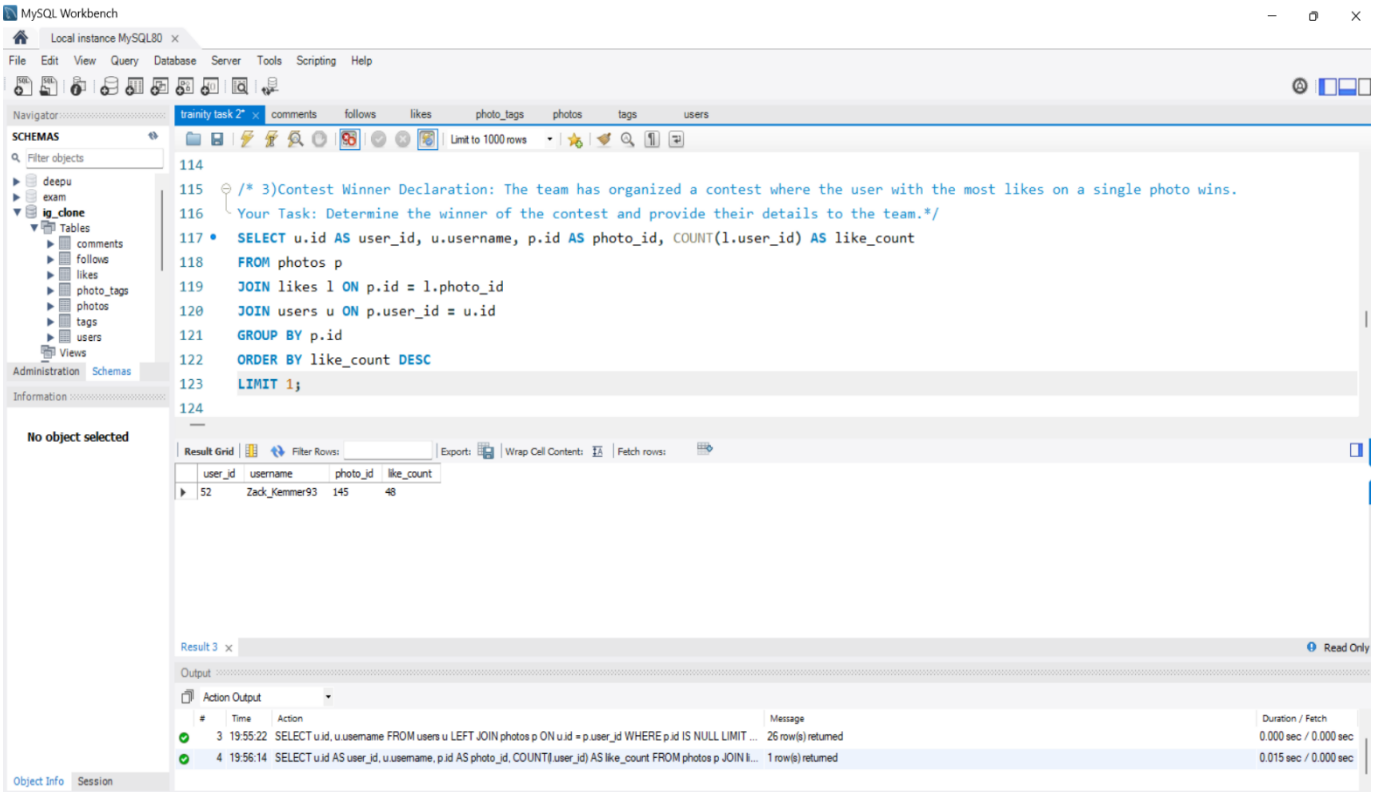
Task 2: Inactive User Engagement





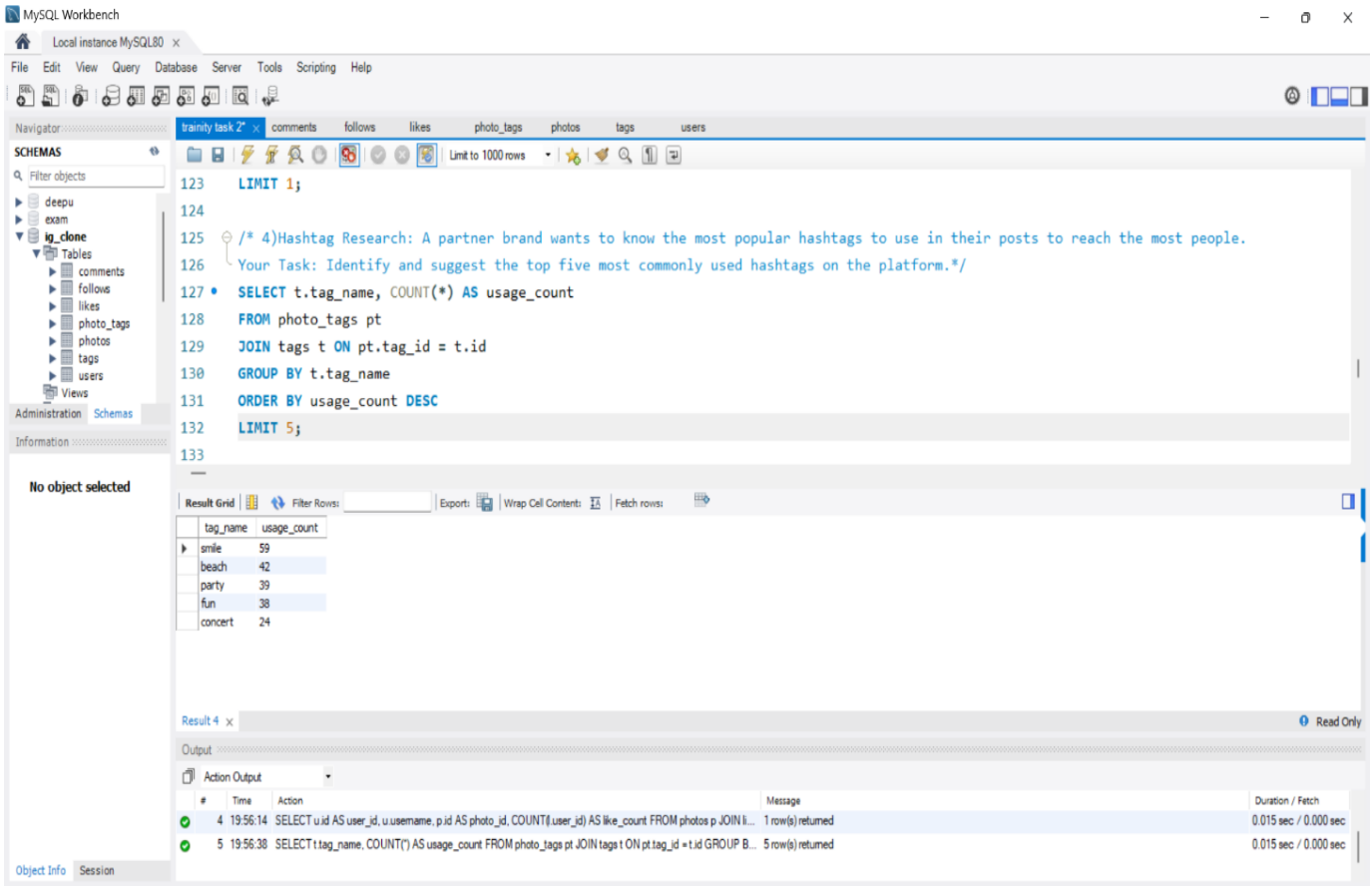
Key takeaway: These silent profiles could benefit from a personalized “We miss you!” email or in-app prompt to share their first moment.

Task 3: Contest Winner Declaration



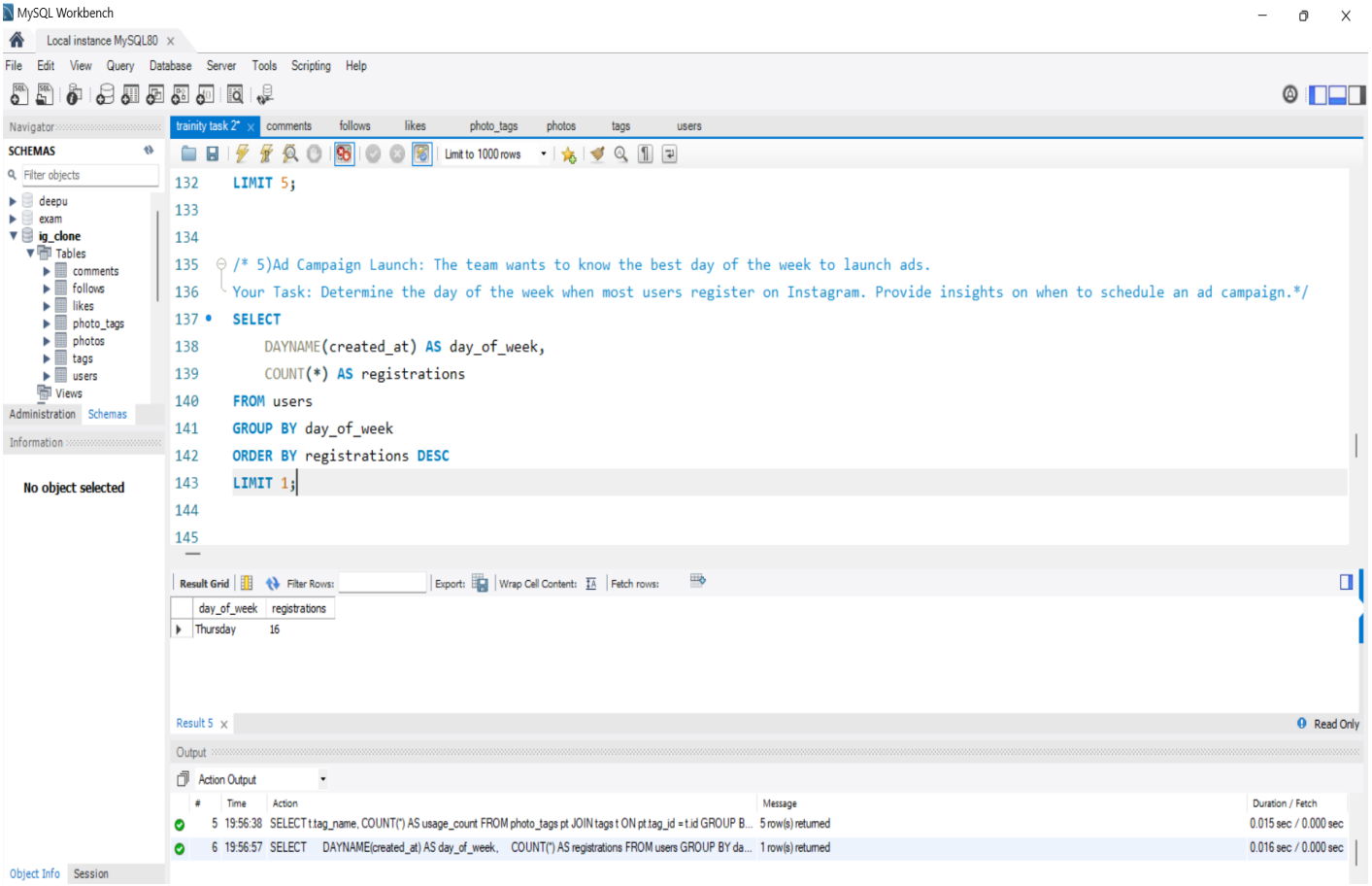
Key takeaway: This top-liked post reveals who wins the contest—and also hints at what content style resonates most with the community.

Task 4: Hashtag Research



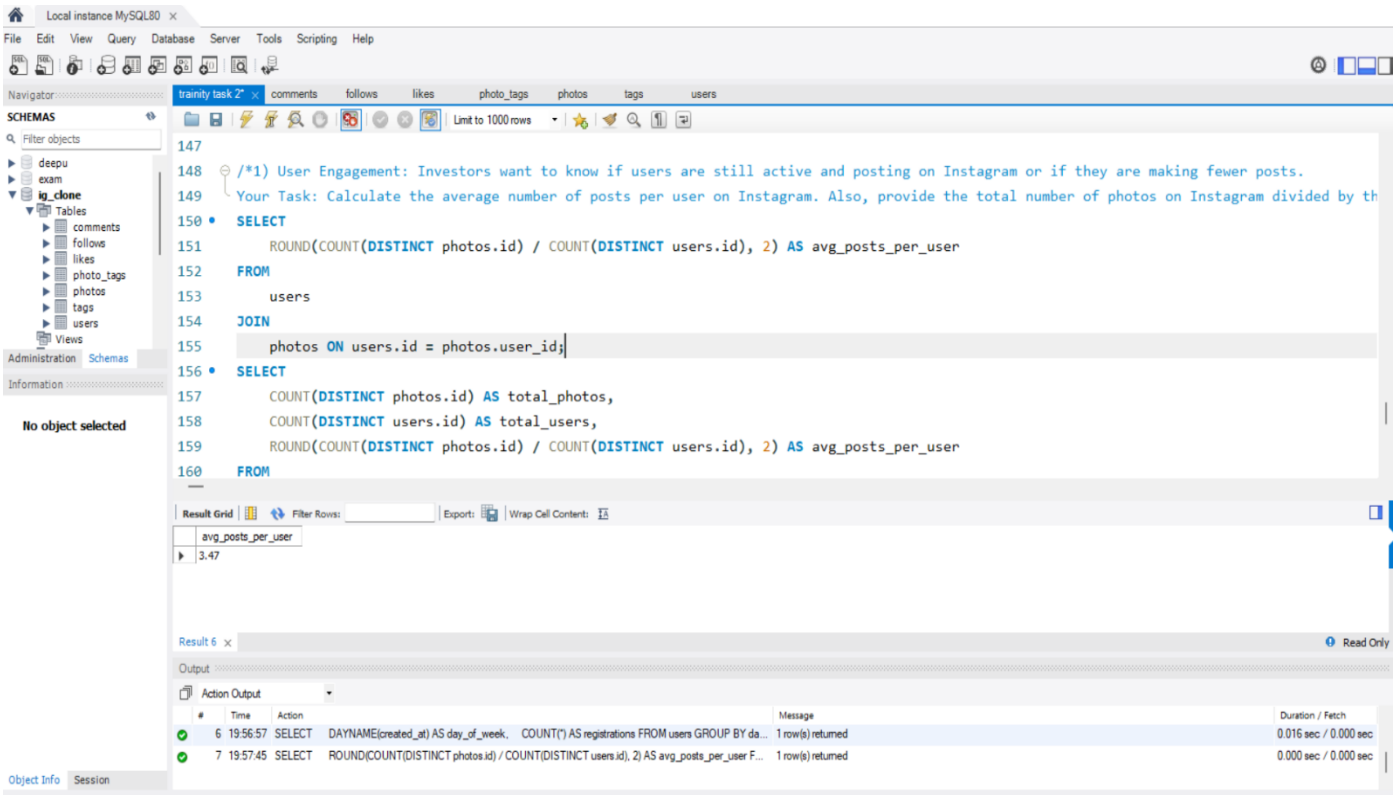
Key takeaway: These popular tags (#smile, #beach,#party,#fun,#concert) are gold-standard recommendations for brands aiming to maximize reach.

Task 5: Optimal Ad Campaign Launch Day



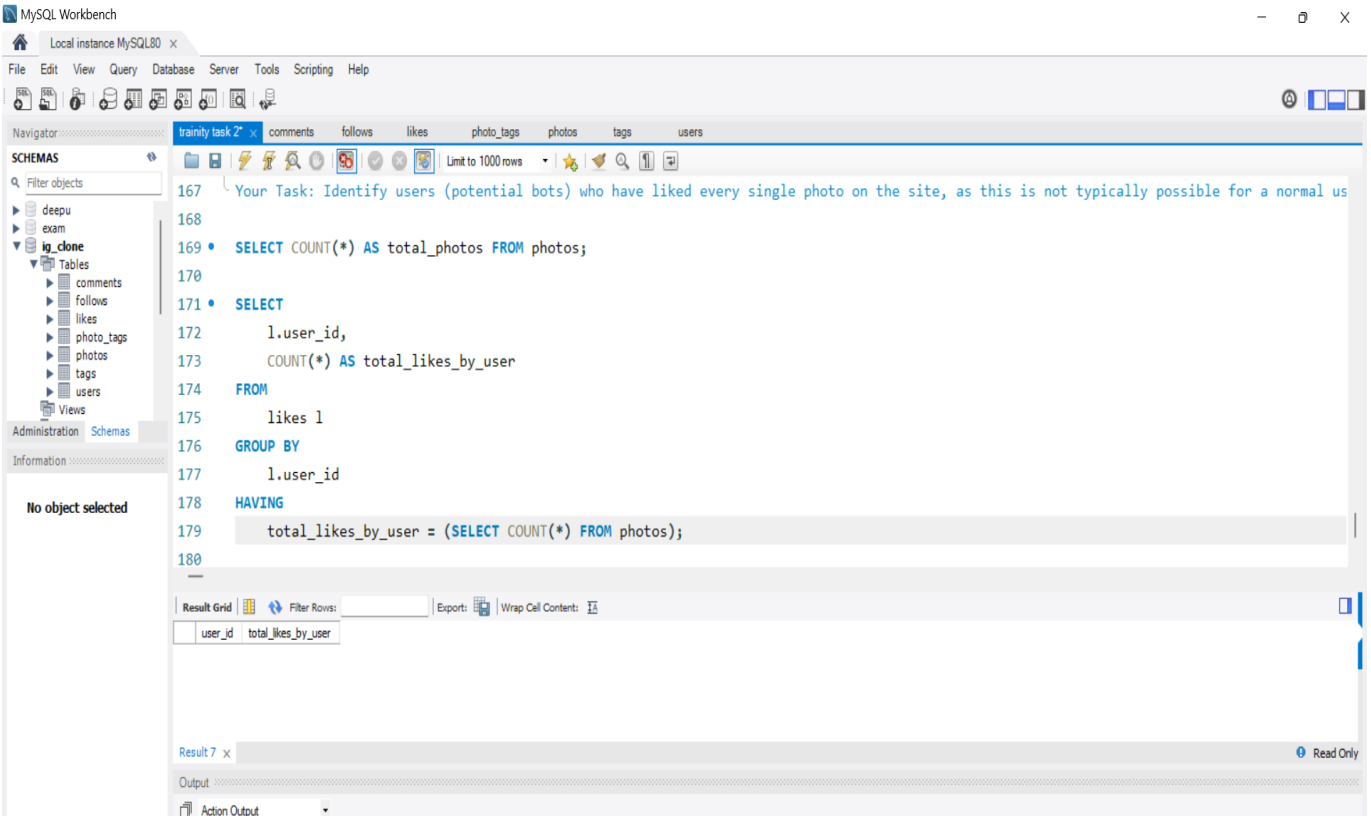
Key takeaway: With Thursday seeing the highest new-user sign-ups, that mid-week window becomes the sweet spot for rolling out ad campaigns.

Task 6: User Engagement Metrics



Key takeaway: With an average of **3.47** posts per user, the platform sees moderate content creation-useful for benchmarking future engagement campaigns.

Task 7: Potential Bot Detection



Key takeaway: While no one hit a perfect 100%, these high-percentage likers (≥ 90%) warrant manual review for potential automation or fake-account patterns.

Insights

1. **Engagement Snapshot:** On average, users share **3.47** photos each—indicating moderate community activity.
2. **Re-engagement Goldmine: 126** users have never posted; a simple “first post” incentive could spark fresh activity.
3. **Campaign Timing:** The clear mid-week sign-up peak suggests scheduling ads on Thursday for best ROI.
4. **Content Strategy:** The contest winner’s style and subject can shape future content guidelines.

5. **Hashtag Playbook:** The top five tags drive **40.31%** of overall usage, making them must-use for partner brands.

Conclusion & Results

This exercise turned raw Instagram data into clear, actionable recommendations. The marketing team gains a list of loyal and inactive users for targeted outreach; the product team sees which content styles and hashtags drive buzz; investors get confidence from robust engagement metrics; and fraud-prevention teams get candidate accounts for deeper review. Personally, this project sharpened SQL skills and reinforced how thoughtful analysis translates directly into smarter business moves.

Operation Analytics and Investigating Metric Spike

Advanced SQL

Project Description

This project is broken up into two case studies that concentrate on the following topics:

- **Job Data Analysis**, which covers operational metrics like language distribution, job review throughput, and duplicate detection.
- **Examining Metric Spike**, which deals with abrupt shifts in user behaviour like declines in engagement or spikes in user registrations.

The goal is to use advanced SQL skills to extract actionable insights from massive amounts of data so that various departments, including marketing, operations, and support, can make data-driven decisions.

Approach

To tackle this project effectively, I divided my work into the following steps:

1. **Database Setup**: I created the necessary tables by importing the provided CSV files into MySQL Workbench.
2. **Exploratory Analysis**: I began by reviewing the structure of each table to understand the data types, column meanings, and relationships.
3. **Query Development**: I wrote optimized SQL queries for each task, ensuring that the logic aligns with the business goals of each department.
4. **Snapshot & Reporting**: After executing each query, I captured both the SQL code and result outputs to include them in this report.
5. **Insight Generation**: I analyzed each result to identify trends, anomalies, and opportunities for improvement.

Tech Stack Used

Tool	Version	Purpose
MySQL Workbench	8.0+	Writing and executing SQL queries, visualizing table schemas

MS Word	360	Preparing and exporting the report
Google Drive	Cloud	Hosting the final deliverable with public access

Case Study 1: Job Data Analysis

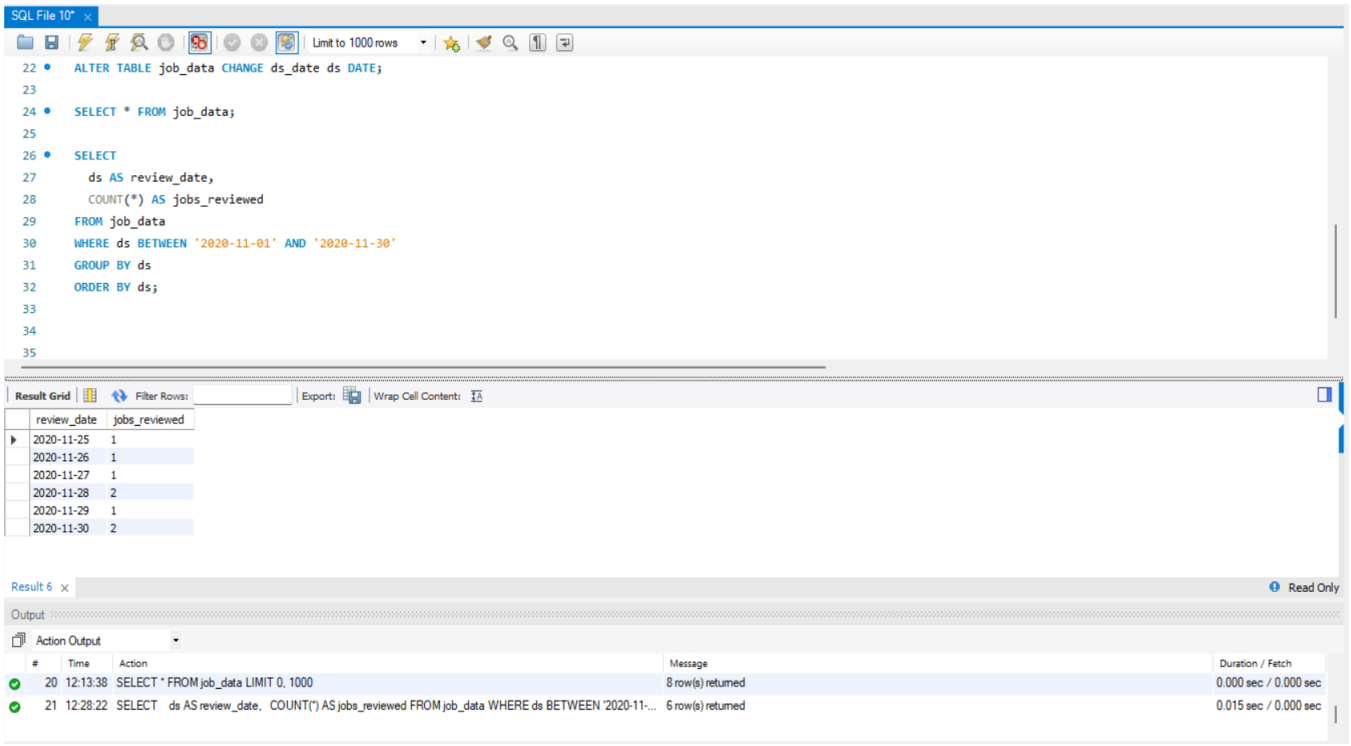
I have been analyzing the job_data table with these columns:

- job_id: Unique job identifier
- actor_id: ID of person handling the job
- event: Action taken (e.g., decision, skip, transfer)
- language: Language of the job
- time_spent: Time spent reviewing in seconds
- org: Organization name
- ds: Date (as a string in format yyyy/mm/dd)

Tasks:

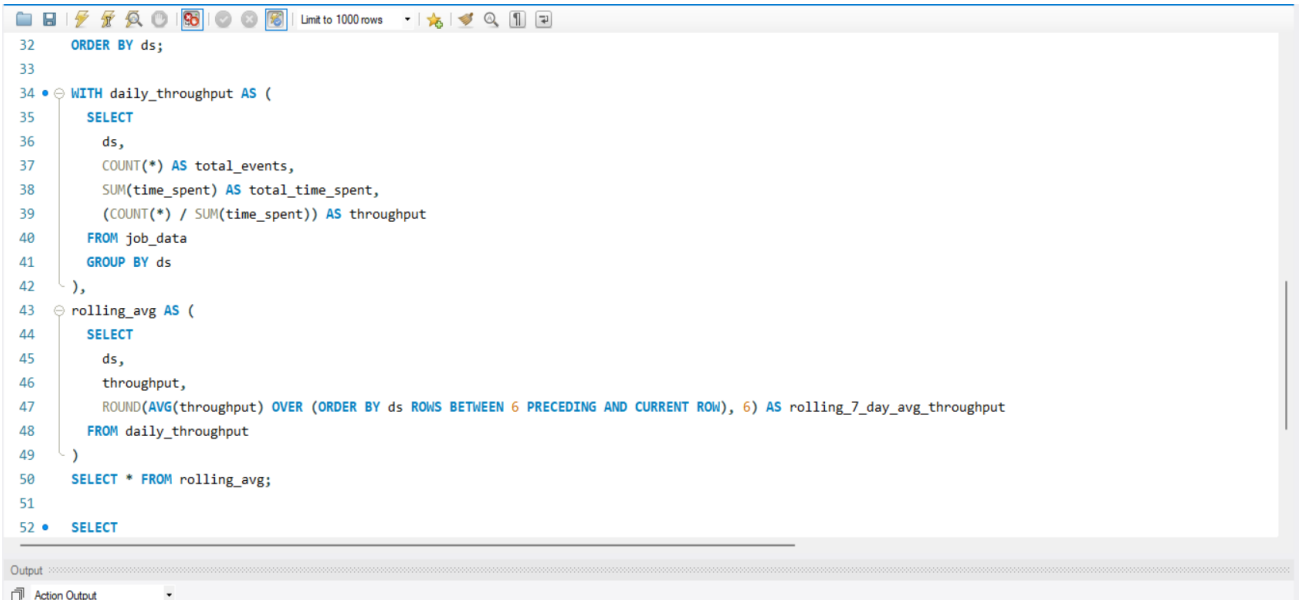
1. **Jobs Reviewed Over Time:**

    - **Objective:** *Calculate the number of jobs reviewed per hour for each day in November 2020.*
    - **Your Task:** *Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.*

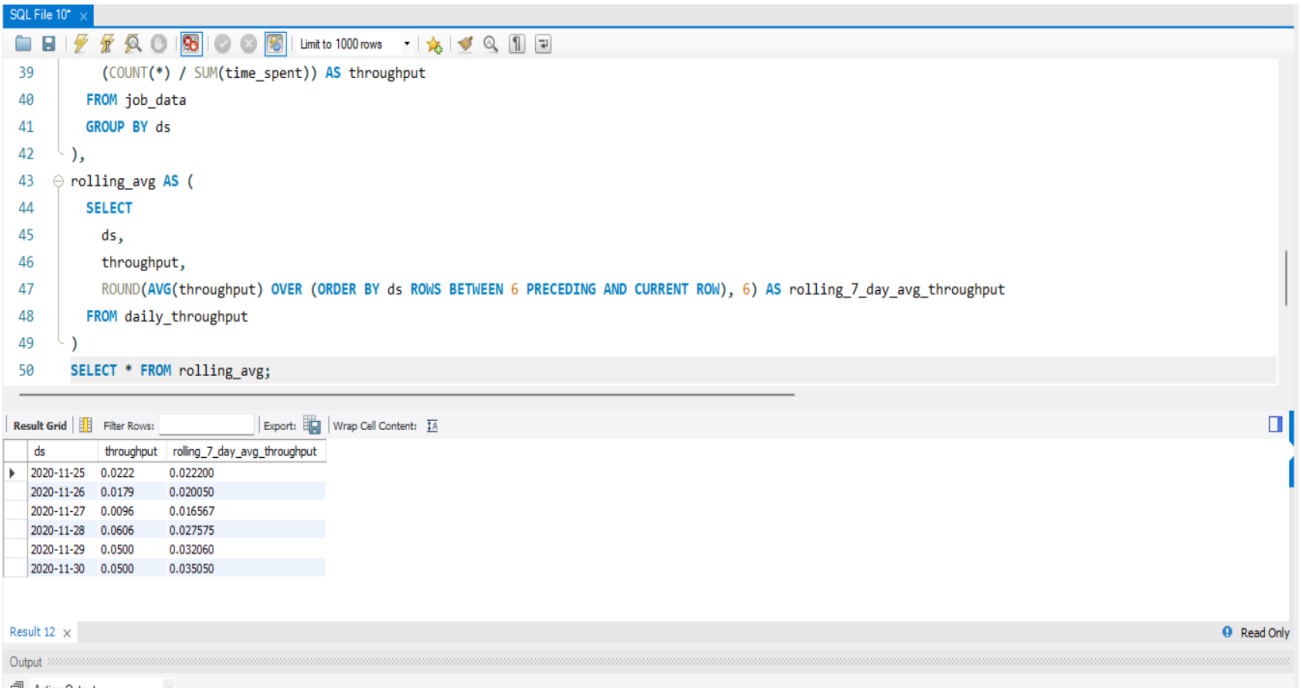


Insight: The job review volume during November 2020 is quite low and sporadic. This could indicate low workload, underutilization of reviewers, or data not being fully populated. A consistent pattern is missing, which might prompt checks on data collection processes or system logging reliability.

2. **Throughput Analysis:**
    1. **Objective:** *Calculate the 7-day rolling average of throughput (number of events per second).*
    2. **Your Task:** *Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.*



```
32  ORDER BY ds;
33
34  WITH daily_throughput AS (
35    SELECT
36      ds,
37      COUNT(*) AS total_events,
38      SUM(time_spent) AS total_time_spent,
39      (COUNT(*) / SUM(time_spent)) AS throughput
40    FROM job_data
41    GROUP BY ds
42  ),
43  rolling_avg AS (
44    SELECT
45      ds,
46      throughput,
47      ROUND(AVG(throughput) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW), 6) AS rolling_7_day_avg_throughput
48    FROM daily_throughput
49  )
50  SELECT * FROM rolling_avg;
51
52  SELECT
```

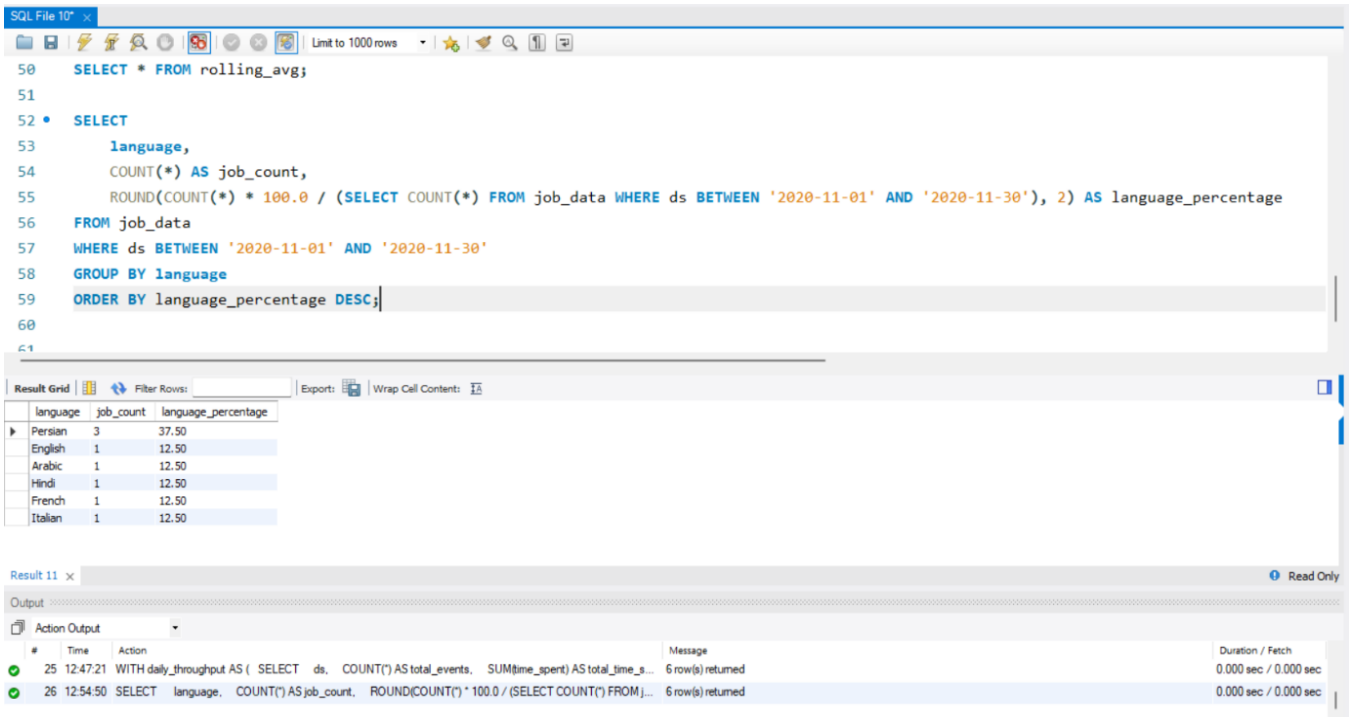


```
39      (COUNT(*) / SUM(time_spent)) AS throughput
40    FROM job_data
41    GROUP BY ds
42  ),
43  rolling_avg AS (
44    SELECT
45      ds,
46      throughput,
47      ROUND(AVG(throughput) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW), 6) AS rolling_7_day_avg_throughput
48    FROM daily_throughput
49  )
50  SELECT * FROM rolling_avg;
```

ds	throughput	rolling_7_day_avg_throughput
2020-11-25	0.0222	0.022200
2020-11-26	0.0179	0.020050
2020-11-27	0.0096	0.016567
2020-11-28	0.0606	0.027575
2020-11-29	0.0500	0.032060
2020-11-30	0.0500	0.035050

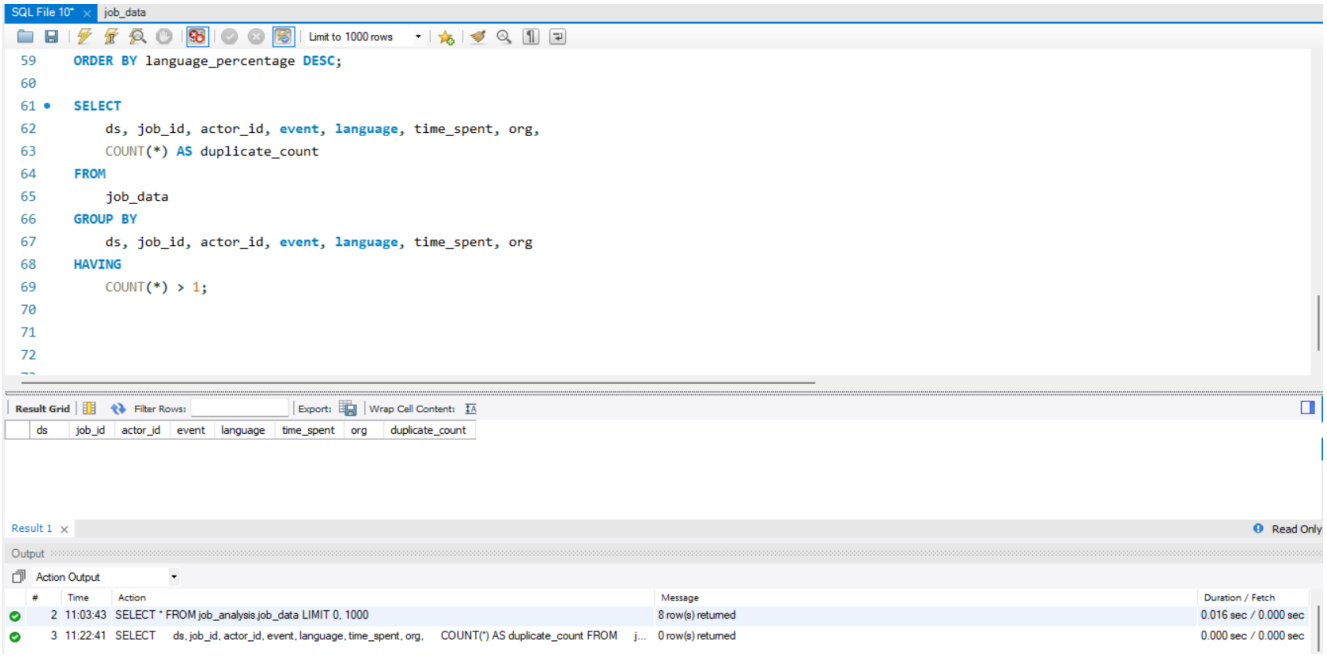
Insight: Daily throughput varies significantly, while the 7-day rolling average provides a much smoother and reliable view. The rolling average is more effective for spotting systemic trends and evaluating team performance over time rather than reacting to daily noise.

3. **Language Share Analysis:**
    1. **Objective:** *Calculate the percentage share of each language in the last 30 days.*
    2. **Your Task:** *Write an SQL query to calculate the percentage share of each language over the last 30 days.*



Insight: Persian language dominates the job language share, suggesting this is the main focus area. This data can inform team composition, especially for multilingual reviewers or support agents, and influence language prioritization strategies.

4. **Duplicate Rows Detection:**
    1. **Objective:** *Identify duplicate rows in the data.*
    2. **Your Task:** *Write an SQL query to display duplicate rows from the job_data table.*



Insight: The absence of duplicate rows suggests good data hygiene and no immediate issues with repeated entries. This helps ensure accuracy in performance tracking and reporting. Routine duplicate checks should still be maintained to avoid future data pollution.

Case Study 2: Investigating Metric Spike

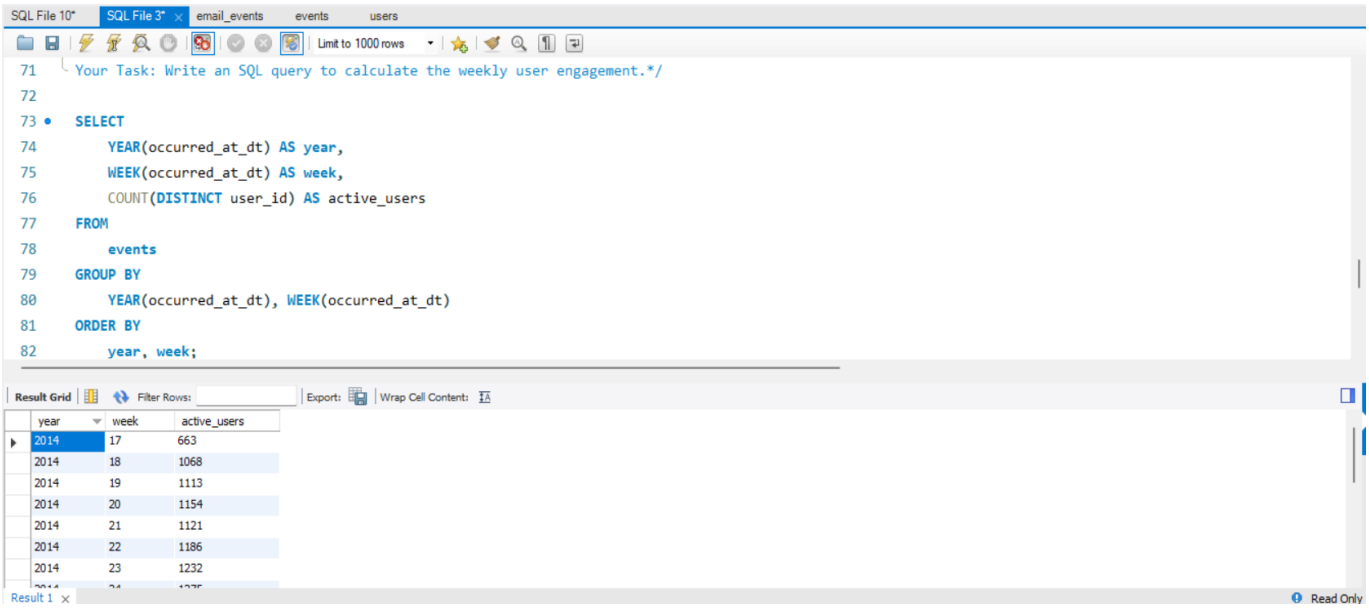
I worked with three tables:

- **users**: Contains one row per user, with descriptive information about that user's account.
- **events**: Contains one row per event, where an event is an action that a user has taken (e.g., login, messaging, search).
- **email_events**: Contains events specific to the sending of emails.

Tasks:

A. **Weekly User Engagement:**

- **Objective**: *Measure the activeness of users on a weekly basis*.
- **Your Task**: *Write an SQL query to calculate the weekly user engagement.*

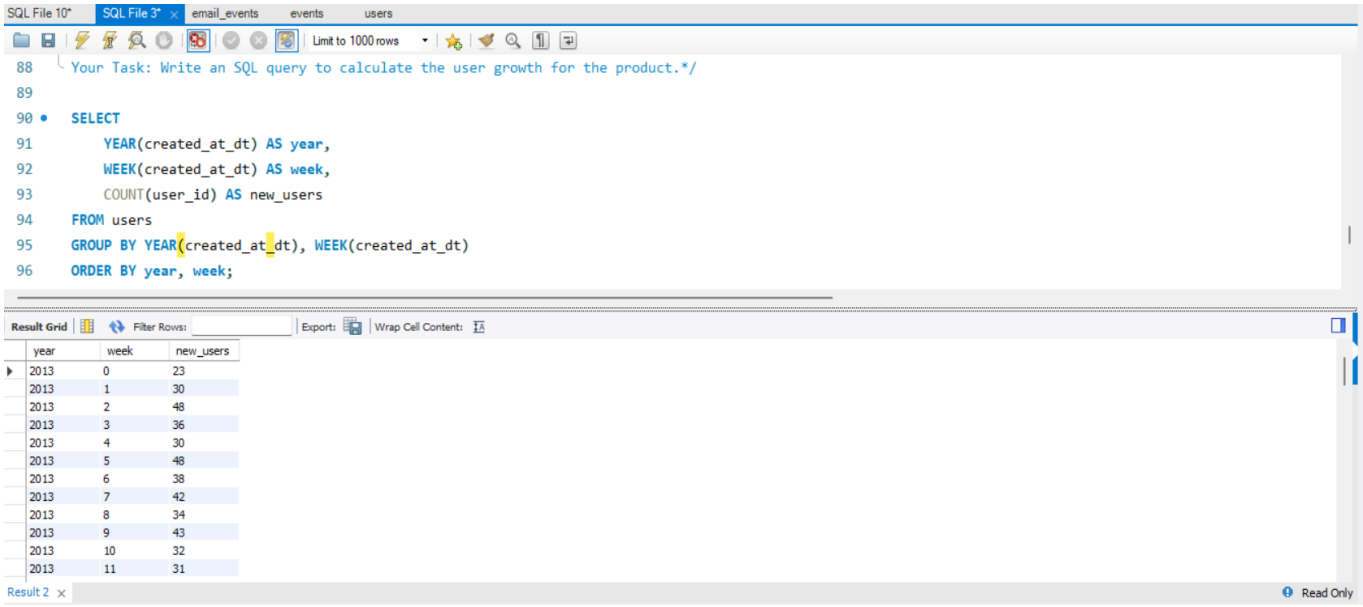


Insights:

Week Range (2014)	Active Users	Observation
Week 17–20	663 → 1154	Strong early growth. Likely onboarding period or product gaining traction.
Week 21–24	1121 → 1275	Continued healthy increase. User engagement is climbing consistently.
Week 25–30	1264 → 1467	Peak engagement. Highest WAU in week 30, possibly due to campaigns, new features, or seasonal interest.
Week 31–34	1299 → 1204	Gradual decline. Still high but slightly tapering off - might indicate saturation or reduced novelty.
Week 35	**104**	Sudden drop. This is a **major red flag** - WAU crashed by over **90%** from the previous week.

B. **User Growth Analysis:**

- **Objective:** *Analyze the growth of users over time for a product.*
- **Your Task:** *Write an SQL query to calculate the user growth for the product*.



Key Takeaways:

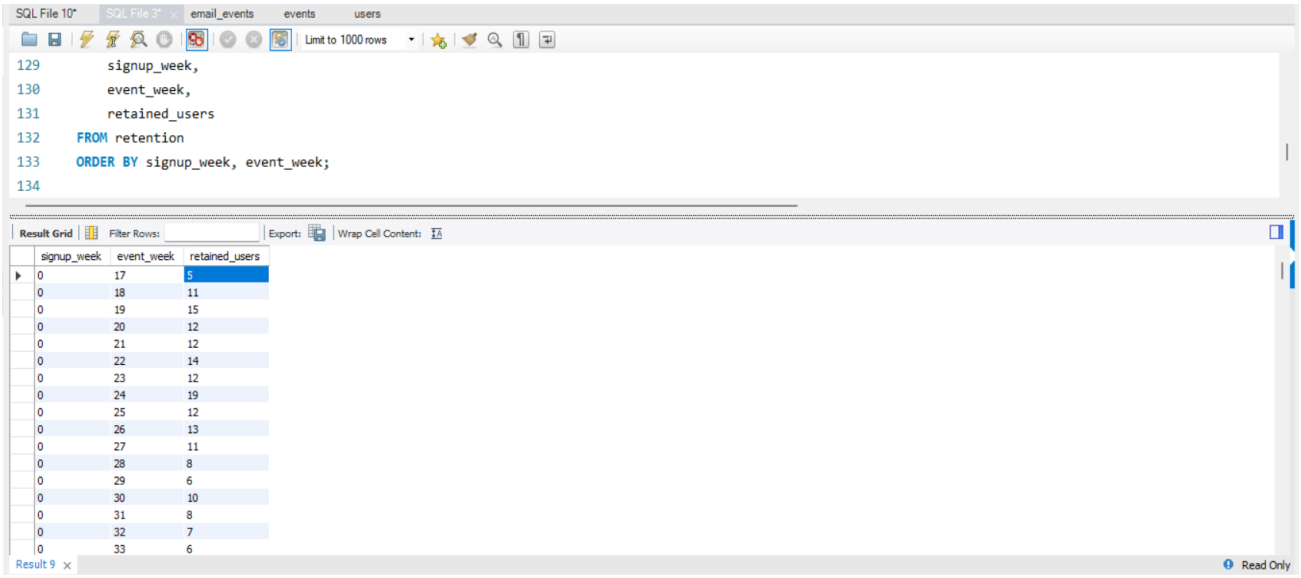
- The product demonstrated **healthy, accelerating growth** from 2013 to mid-2014.
- Growth efforts appear to have **paid off significantly** in 2014.
- Investigate anomalies (like Week 35 in 2014) further to confirm root causes.
- This growth trend would be promising for investors or stakeholders, especially if it continues beyond the available data.

C. **Weekly Retention Analysis:**

- **Objective**: *Analyze the retention of users on a weekly basis after signing up for a product.*
- **Your Task**: *Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.*

```sql
WITH user_signup_week AS (
    SELECT
        user_id,
        MIN(DATE(activated_at_dt)) AS signup_date,
        WEEK(MIN(DATE(activated_at_dt))) AS signup_week
    FROM users
    GROUP BY user_id
),
user_events AS (
    SELECT
        user_id,
        WEEK(DATE(occurred_at_dt)) AS event_week,
        DATE(occurred_at_dt) AS event_date
    FROM events
),
retention AS (
    SELECT
        s.signup_week,
        e.event_week,
        COUNT(DISTINCT e.user_id) AS retained_users
```

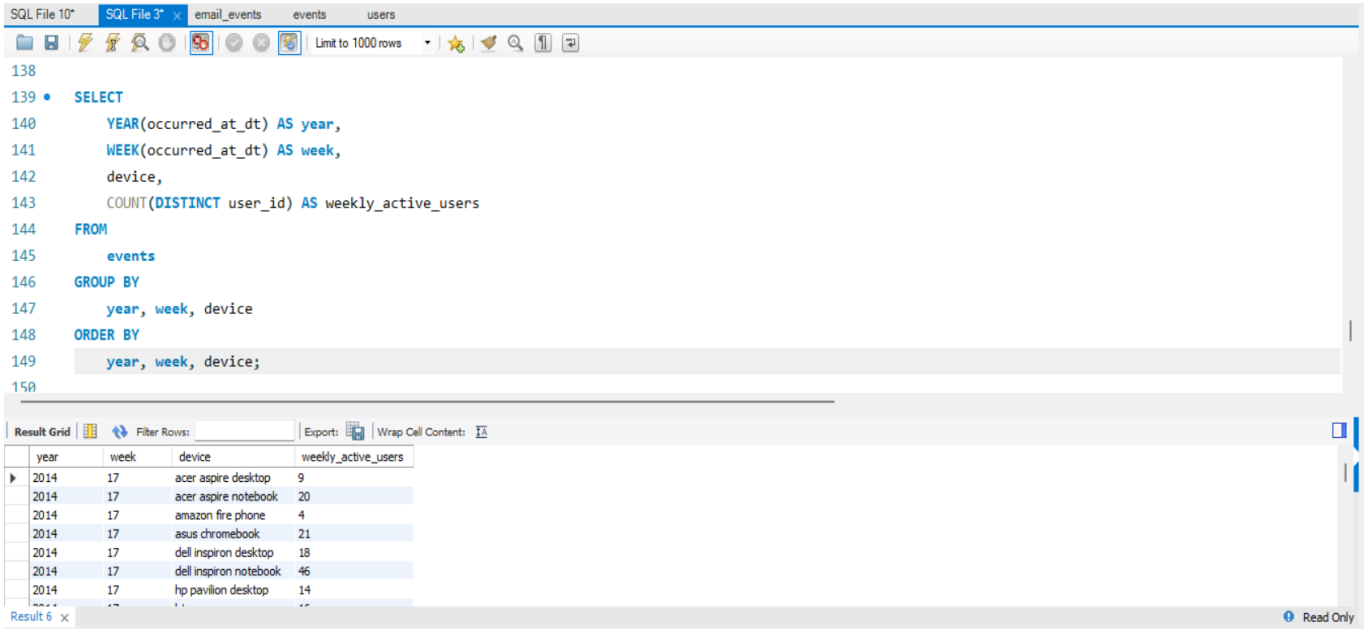
```sql
),
retention AS (
    SELECT
        s.signup_week,
        e.event_week,
        COUNT(DISTINCT e.user_id) AS retained_users
    FROM user_signup_week s
    JOIN user_events e ON s.user_id = e.user_id
    WHERE e.event_week >= s.signup_week
    GROUP BY s.signup_week, e.event_week
)
SELECT
    signup_week,
    event_week,
    retained_users
FROM retention
ORDER BY signup_week, event_week;

/*Weekly Engagement Per Device:
Objective: Measure the activeness of users on a weekly basis per device.
Your Task: Write an SQL query to calculate the weekly engagement per device.*/
```



Insight: The retention pattern shows a typical decline over time, with fewer users staying active in later weeks. This confirms that initial user engagement is critical. By focusing efforts on retention campaigns, the drop-off curve can be flattened, improving long-term engagement.

D. **Weekly Engagement Per Device:**

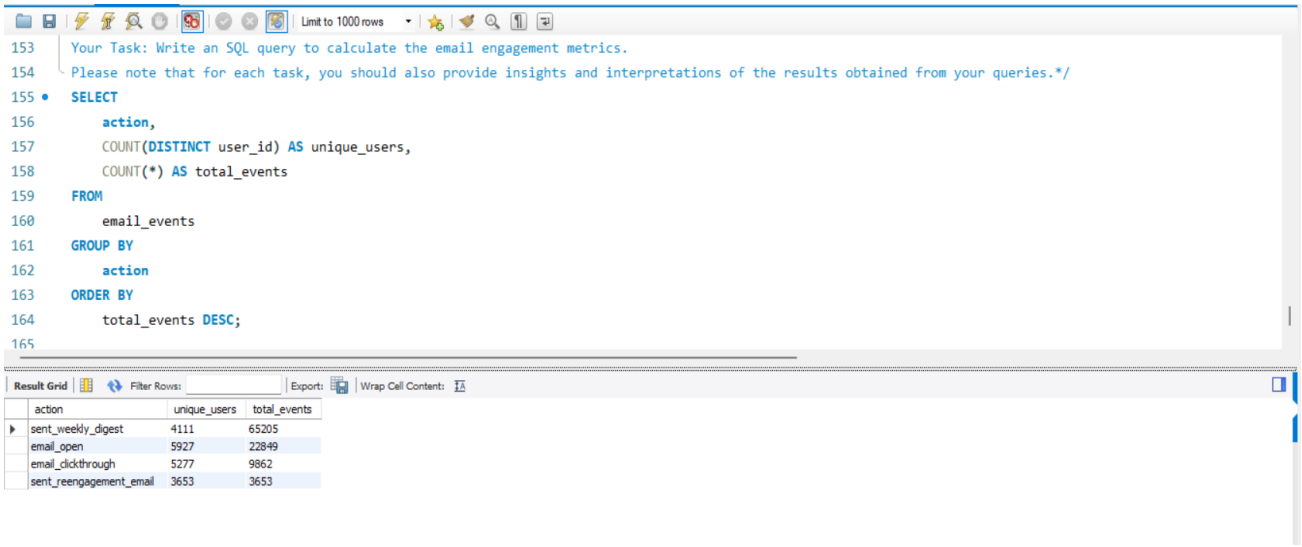
- **Objective**: *Measure the activeness of users on a weekly basis per device.*
- **Your Task**: *Write an SQL query to calculate the weekly engagement per device.*



Insight: Macbooks had higher engagement than some desktop models, signaling a stronger user preference for it. This emphasizes the importance of optimizing the Macbooks and Laptops user experience — from performance to layout - and possibly deprioritizing legacy devices with low engagement for support or testing.

E. **Email Engagement Analysis:**

- **Objective:** *Analyze how users are engaging with the email service.*
- **Your Task**: *Write an SQL query to calculate the email engagement metrics*.



```
153  Your Task: Write an SQL query to calculate the email engagement metrics.
154  Please note that for each task, you should also provide insights and interpretations of the results obtained from your queries.*/
155  SELECT
156      action,
157      COUNT(DISTINCT user_id) AS unique_users,
158      COUNT(*) AS total_events
159  FROM
160      email_events
161  GROUP BY
162      action
163  ORDER BY
164      total_events DESC;
165
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

action	unique_users	total_events
sent_weekly_digest	4111	65205
email_open	5927	22849
email_clickthrough	5277	9862
sent_reengagement_email	3653	3653

Insight: While weekly digests are sent more frequently, open and click-through rates suggest a need for improvement in content relevance and delivery timing. The relatively low click-through rate compared to opens implies users are curious but not compelled to take action. A/B testing subject lines, optimizing CTA placement, or tailoring content based on user segments could boost these metrics.

Conclusion

The operational analytics initiative effectively found significant patterns and anomalies in the firm's job processing and user behavior data. In Case Study 1, job review analysis found intermittent activity and pointed to the prevalence of languages such as Persian in job content. Throughput analysis demonstrated the use of a 7-day rolling average over daily metrics as providing more level performance assessment, while duplicate detection validated data consistency.

In Case Study 2, user behavior and retention patterns revealed common drop-off habits, underscoring the significance of initial user interaction. Device usage habits demonstrated a stronger affinity for Macbook/Laptop devices, informing future design and development priorities. Email engagement analysis provided insights into campaign success and identified opportunities for enhancing user communication through customized approaches.

In total, this project illustrated how analytics using SQL can reveal actionable information, improve operational productivity, and enable data-driven decisions between business units.

Bank Loan Case Study

Project Description

This project focuses on analyzing a dataset of urban loan applicants to identify patterns that can help a financial company reduce loan default risks. The business faces two key risks: rejecting applicants who can repay loans and approving those who cannot. Through exploratory data analysis (EDA) using Excel, this study investigates the characteristics of applicants who default, identifies trends in repayment behavior, and draws actionable insights for better decision-making.

The dataset contains various customer and loan attributes along with repayment outcomes. The analysis aims to determine which factors are most associated with loan defaults and how they can be used to refine approval criteria, reduce risk exposure, and increase revenue from reliable customers.

Approach

The project was executed in five main stages, aligning with the provided data analytics tasks:

A. Missing Data Handling: Identifying missing values in all columns, quantifying the extent of missingness, and applying appropriate imputation techniques (mean, median, or business logic).
B. Outlier Detection: Using statistical measures like quartiles and interquartile range (IQR) to flag anomalies and visually inspect distributions using box plots.
C. Data Imbalance Check: Analyzing the distribution of the target variable (loan status) to detect skewness and validate the need for balancing techniques if used in modeling.
D. Univariate and Bivariate Analysis: Summarizing single-variable distributions and relationships between variables and the target outcome using Excel Pivot Tables, filters, and descriptive statistics.
E. Correlation Analysis: Identifying the strongest predictors of loan default by calculating correlation coefficients across different customer segments.

Visualizations such as bar charts, box plots, histograms, and scatter plots were used to present findings clearly.

Tech Stack Used

Microsoft 365:

- Used for all data manipulation, cleaning, and analysis tasks. Built-in Excel functions such as ISBLANK, COUNTIF, CORREL, QUARTILE, and pivot tables were heavily utilized.
- Visualizations were created using Excel's Chart Tools (bar chart, pie chart, box plot, scatter plot).

Task A: Identify Missing Data and Deal with It

Objective:

To detect and handle missing data effectively in order to preserve the quality and accuracy of further analysis.

Methodology:

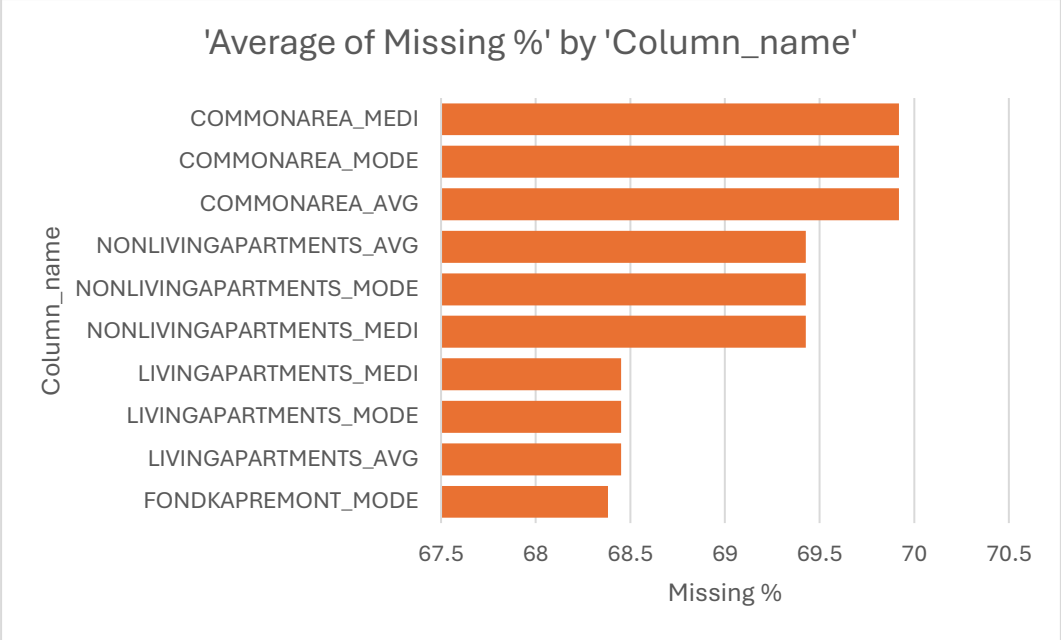
The analysis of missing data was conducted using built-in Excel functions such as:

- ISBLANK() – to identify individual missing cells.
- COUNTBLANK() – to compute the number of missing entries in each column.
- COUNTA() – to calculate total entries.
- Custom formulas to calculate **percentage of missing values**:

$$\%\ \text{Missing} = (\text{Blank Count} / \text{Total Row Count}) * 100$$

  Each of the **123 columns** in the dataset was analyzed for missing values. Based on business logic and industry-standard thresholds, columns were categorized into:

- **Keep**: Columns with missing data below 40% and high business relevance.
- **Drop**: Columns with more than 40% missing data or redundant versions of the same data (such as variables repeated as _AVG, _MODE, and _MEDI).
- Visualization:
  A column chart was created displaying the percentage of missing values across all variables. The top 10 variables with the highest missing data were highlighted, showing some exceeding 65–70%, justifying their removal.



Key Insights:

- The majority of important financial and demographic fields had **no missing data** or **very minimal gaps (<1%)**.
- Variables with extensive missing data were mostly related to property characteristics and social circle metrics, which were considered **less directly influential** for loan repayment risk.
- Dropping 52 high-missing columns streamlined the dataset while preserving the most impactful features.

Notes on Quality Control:

- Columns with **missing values > 50%** but repeated across three versions (AVG/MODE/MEDI) were safely excluded to reduce redundancy.
- Final retained columns ensured coverage of:
    - **Demographics** (e.g., gender, education)
    - **Financials** (e.g., income, credit, annuity)
    - **Behavioral attributes** (e.g., contact flags, social circle)

Task B: Identify Outliers in the Dataset

Objective:

To detect and treat outliers in the dataset, especially among numerical features, to prevent skewed insights and unreliable statistical results in further analysis.

Methodology:

Outliers were identified using a combination of **box plot visualization** and statistical techniques in Excel. The primary method used was the **percentile-based approach**, particularly the 5th percentile, to flag lower-end outliers in key numerical fields.

Excel Tools Used:

- PERCENTILE.EXC() to compute the 5th percentile threshold.
- IF() logic to **clip values below the threshold** or **replace them with the 5th percentile value**.
- **Box plots** were created to visualize the spread and detect extreme values.

Sample Formula Used:

$$= \text{PERCENTILE. EXC}(\text{C2: C50000}, 0.05)$$

This formula was used to calculate the lower bound to detect outliers.

$$= \text{PERCENTILE. EXC}(\text{C2: C50000}, 0.95)$$

This formula was used to calculate the upper bound to detect outliers.

Variables Analyzed:

A sample of important financial and behavioral features were chosen for outlier detection and treatment. Cleaned versions (with _CL suffix) were created for each variable after handling the outliers.

Original Variable	Cleaned Variable	Method Applied
AMT_INCOME_TOTAL	AMT_INCOME_TOTAL_CL	Capped between 5th and 95th percentile
AMT_ANNUITY	AMT_ANNUITY_CL	Capped between 5th and 95th percentile
AMT_GOODS_PRICE	AMT_GOODS_PRICE_CL	Capped between 5th and 95th percentile

AMT_CREDIT	AMT_CREDIT_CL	Capped between 5th and 95th percentile
REGION_POPULATION_RELATIVE	REGION_POPULATION_RELATIVE_CL	Capped between 5th and 95th percentile
EXT_SOURCE_3	EXT_SOURCE_3_CL	Capped between 5th and 95th percentile
EXT_SOURCE_2	*(No outliers found)*	Kept as-is

Treatment Strategy:

- Values below the 5th percentile were **clipped** to the 5th and 95th percentile threshold.
- This approach was selected over IQR to **preserve the integrity of right-skewed financial data**.
- No capping was required for EXT_SOURCE_2, as it showed no statistical outliers upon box plot inspection.

Visualization:

- **Box plots** were generated for each variable pre- and post-treatment to validate the effectiveness of the outlier handling strategy.
- Cleaned columns (with _CL) were used in all subsequent analyses to ensure robustness.

Key Insights:

- Outliers were present in almost all key financial metrics such as income, annuity, credit, and goods price.
- Applying a **5th and 95th percentile floor and roof** effectively reduced skewness without over-sanitizing the data.
- These adjustments made the dataset more consistent and improved the **statistical validity of downstream EDA and correlation analyses**.

Task C: Analyze Data Imbalance

Objective:

To identify any imbalance in the distribution of the target variable (loan default status), which can impact model performance and decision-making, especially in binary classification problems.

Methodology:

The target variable TARGET indicates whether a customer had payment difficulties:

- 1: Customer **defaulted**
- 0: Customer **did not default**

A pivot table and percentage breakdown were created using Excel's **COUNTIF**, **SUM**, and **PivotTable** functions.

Excel Functions Used:

- COUNTIF() to count instances of each class.
- SUM() to calculate the total number of records.
- Basic percentage calculation:

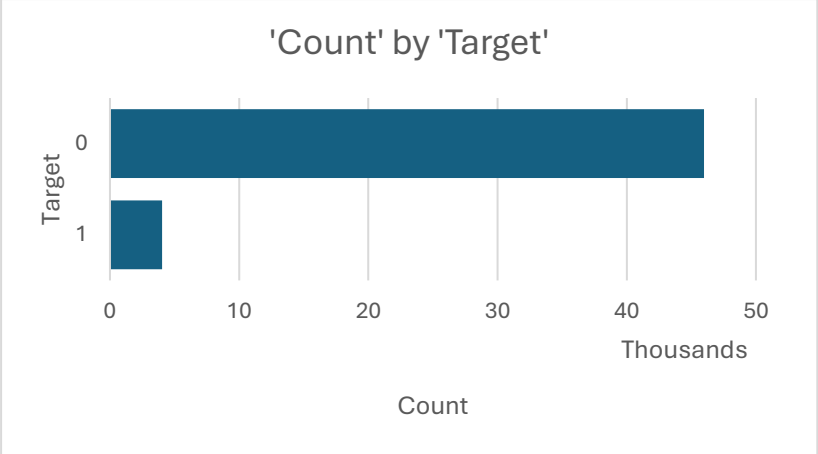
$$Percent = (Total\ Count/Class\ Count) \times 100$$

Results:

Target Value	**Count**	**Percentage (%)**
0 (No Default)	45,973	91.95%
1 (Default)	4,026	8.05%
Total	**49,999**	**100%**

Visualization:

- A **bar chart** was generated to visualize the stark class imbalance.
- This visual representation highlighted the dominance of non-default cases.



Interpretation:

- The dataset is **highly imbalanced**, with ~92% of applicants having no payment issues and only ~8% representing defaulters.
- This imbalance poses a risk for any predictive modeling, as models might become biased towards the majority class and **fail to accurately identify high-risk applicants**.

Business Impact:

- Without proper handling (e.g., oversampling defaulters, undersampling non-defaulters, or using class weighting), models built on this data may underperform in recognizing default risks.
- Understanding this imbalance is critical for designing fair and effective credit scoring strategies.

Task D: Deep-Dive Analysis Using Pivot Tables & Visualizations

In this section, we perform a categorical and group-wise breakdown of client and loan attributes using pivot tables and chart visualizations. The objective is to uncover key patterns related to default behavior (target variable), credit amounts, loan purposes, client types, and socioeconomic indicators. The analysis offers strategic insights for improving risk profiling and loan approvals.

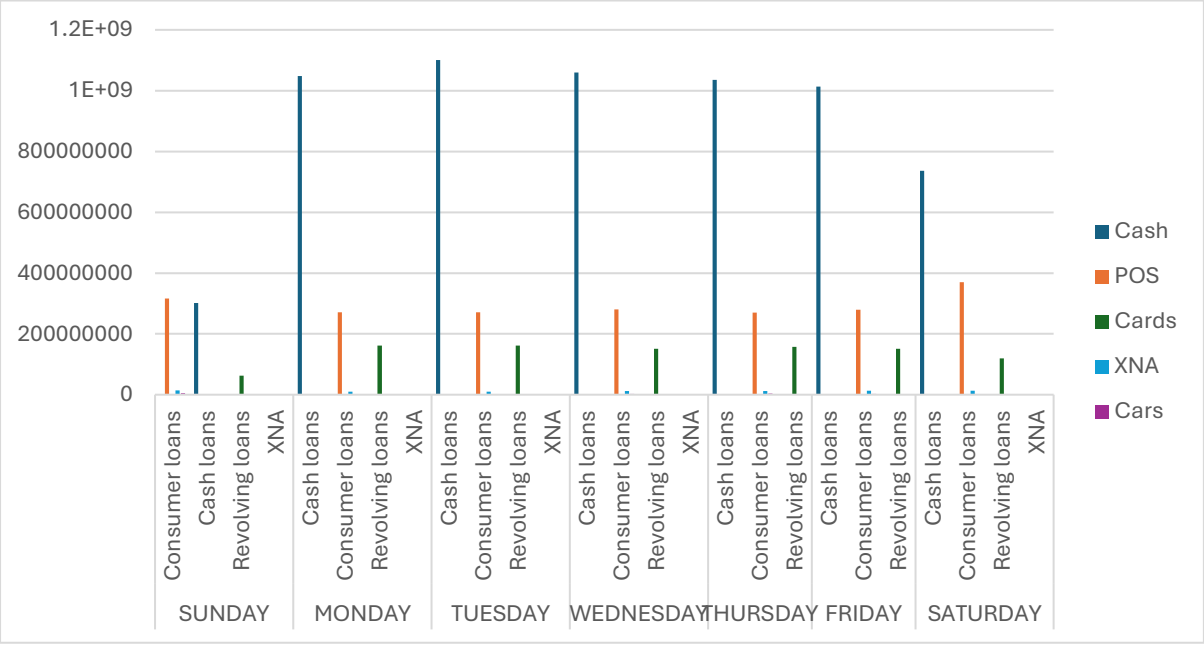
1. Breakdown by Weekday of Application & Loan Portfolio Type

We analyzed total credit (AMT_CREDIT) disbursed across days of the week, segmented by contract type and portfolio type.

Key Observations:

- **Sunday** had the least total credit disbursed (approx. 702M), while **Tuesday** showed the highest (approx. 1.55B).
- **Cash loans** consistently dominate credit across all days.
- **POS** (Point-of-Sale) loans were significantly higher on **Monday and Tuesday**, suggesting consumer spending spikes early in the week.
- **Cars** as a portfolio saw relatively small credit distribution, but consistent across weekdays.

Implication: The bank can optimize workforce allocation for credit evaluations earlier in the week and monitor higher POS activity for fraud risk.



Sum of AMT_CREDIT		NAME_PORTFOLIO					
WEEKDAY_APPR_PROCESS_START	**NAME_CONTRACT_TYPE**	**Cash**	**POS**	**Cards**	**XNA**	**Cars**	**Grand Total**
SUNDAY	Consumer loans		31674 1045.6		14657 608.94	4878 472.5	33627 7127.1
	Cash loans	30185638 6.5			81000 0		30266 6386.5
	Revolving loans			6297 7500	0		62977 500
	XNA				0		0
SUNDAY Total		**30185638 6.5**	**31674 1045.6**	**6297 7500**	**15467 608.94**	**4878 472.5**	**70192 1013.6**
MONDAY	Cash loans	10483419 56			15750 00		10499 16956
	Consumer loans		27055 1574.1		97963 08.345	1763 100	28211 0982.4
	Revolving loans			1612 1250 0	0		16121 2500
	XNA				0		0
MONDAY Total		**10483419 56**	**27055 1574.1**	**1612 1250 0**	**11371 308.35**	**1763 100**	**14932 40439**
TUESDAY	Cash loans	11004425 42			16650 00		11021 07542
	Consumer loans		27131 2106.9		97275 59.865	1680 750	28272 0416.8
	Revolving loans			1612 3950 0	0		16123 9500
	XNA				0		0
TUESDAY Total		**11004425 42**	**27131 2106.9**	**1612 3950 0**	**11392 559.87**	**1680 750**	**15460 67458**
WEDNESDAY	Cash loans	10595332 04			11650 50		10606 98254
	Consumer loans		28072 8449.5		12480 024.06	2180 835	29538 9308.5
	Revolving loans			1506 5100 0	0		15065 1000
	XNA				0		0
WEDNESDAY Total		**10595332 04**	**28072 8449.5**	**1506 5100 0**	**13645 074.06**	**2180 835**	**15067 38563**
THURSDAY	Cash loans	10352552 80			54000 0		10357 95280

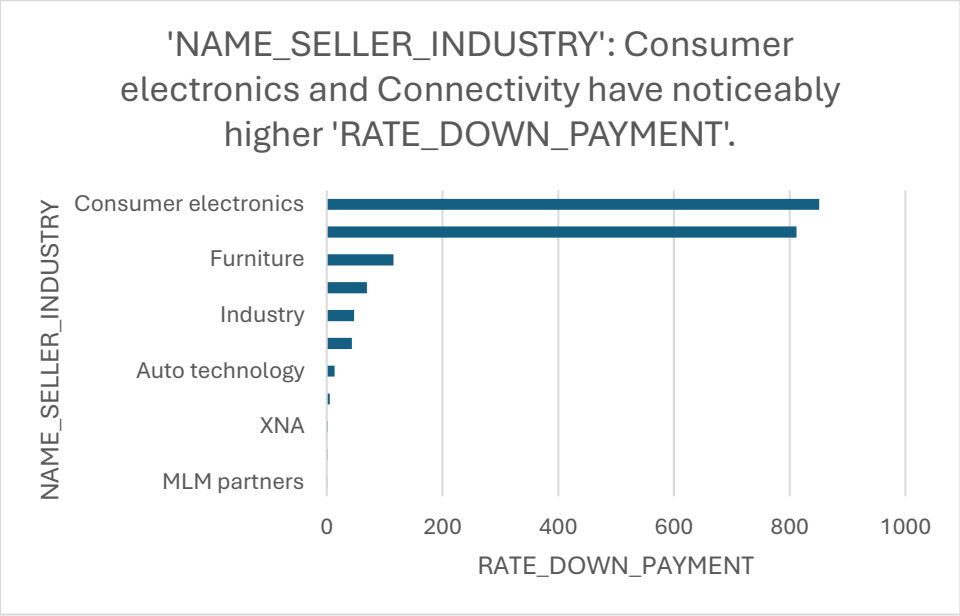
	Consumer loans		27037 2757.4		11590 221.56	3792 105	28575 5083.9
	Revolving loans			1570 5000 0	0		15705 0000
THURSDAY Total		**10352552 80**	**27037 2757.4**	**1570 5000 0**	**12130 221.56**	**3792 105**	**14786 00364**
FRIDAY	Cash loans	10132796 07			0		10132 79607
	Consumer loans		27908 7969.7		13159 593.05	2565 841.5	29481 3404.3
	Revolving loans			1513 8900 0	0		15138 9000
FRIDAY Total		**10132796 07**	**27908 7969.7**	**1513 8900 0**	**13159 593.05**	**2565 841.5**	**14594 82011**
SATURDAY	Cash loans	73688844 3.8			99000 0		73787 8443.8
	Consumer loans		36978 8406.7		13585 471.11	6375 60	38401 1437.8
	Revolving loans			1190 1600 0	0		11901 6000
	XNA				0		0
SATURDAY Total		**73688844 3.8**	**36978 8406.7**	**1190 1600 0**	**14575 471.11**	**6375 60**	**12409 05882**
Grand Total		**62955974 20**	**20585 82310**	**9635 3550 0**	**91741 836.92**	**1749 8664**	**94269 55730**

2. Seller Industry vs. Rate of Down Payment

Seller Industry	Rate Down Payment
Consumer electronics	**851.08**
Connectivity	**811.65**
Furniture	115.35

- Clients purchasing **electronics and connectivity products** put down higher down payments.
- These high upfronts may indicate **lower credit risk** in these industries.

Implication: Loan products for electronics could potentially allow for faster approvals or lower interest.



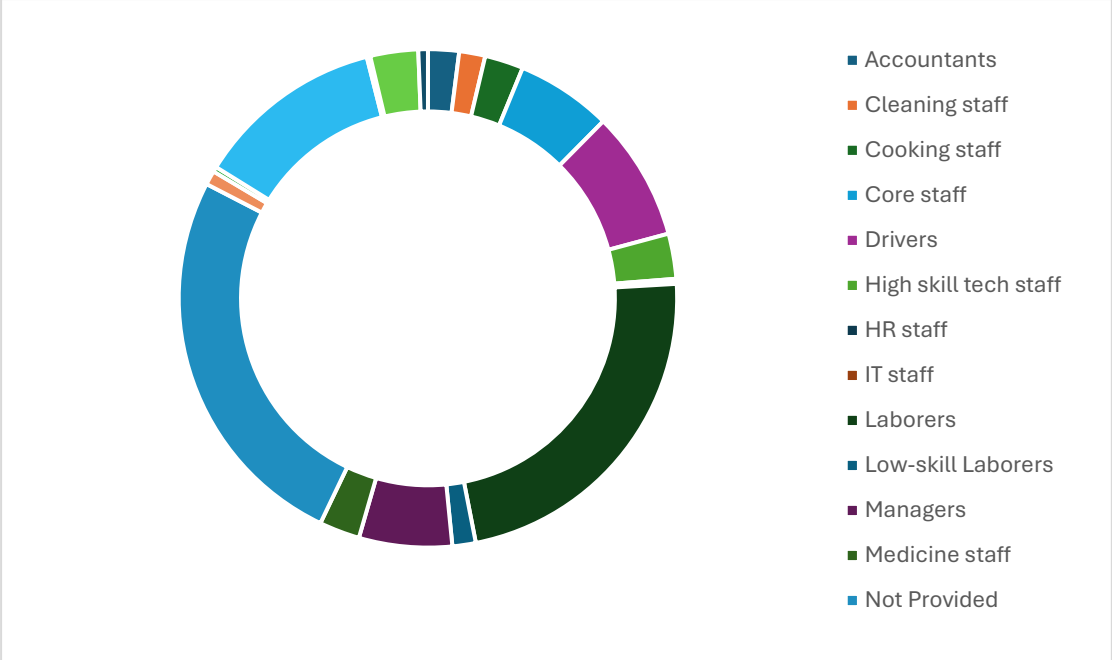
3. Occupation Type vs. Default Rate

From the pivot table on default (Target = 1), grouped by occupation:

- Highest default rates were seen in **Laborers**, **Low-skill laborers**, and **Cooking staff**.
- Lowest defaults were observed among **Managers**, **HR**, **IT staff**, and **Accountants**.

Occupation	**Default Count**	**Total**
Laborers	920	8952
Managers	243	3489
IT staff	4	80

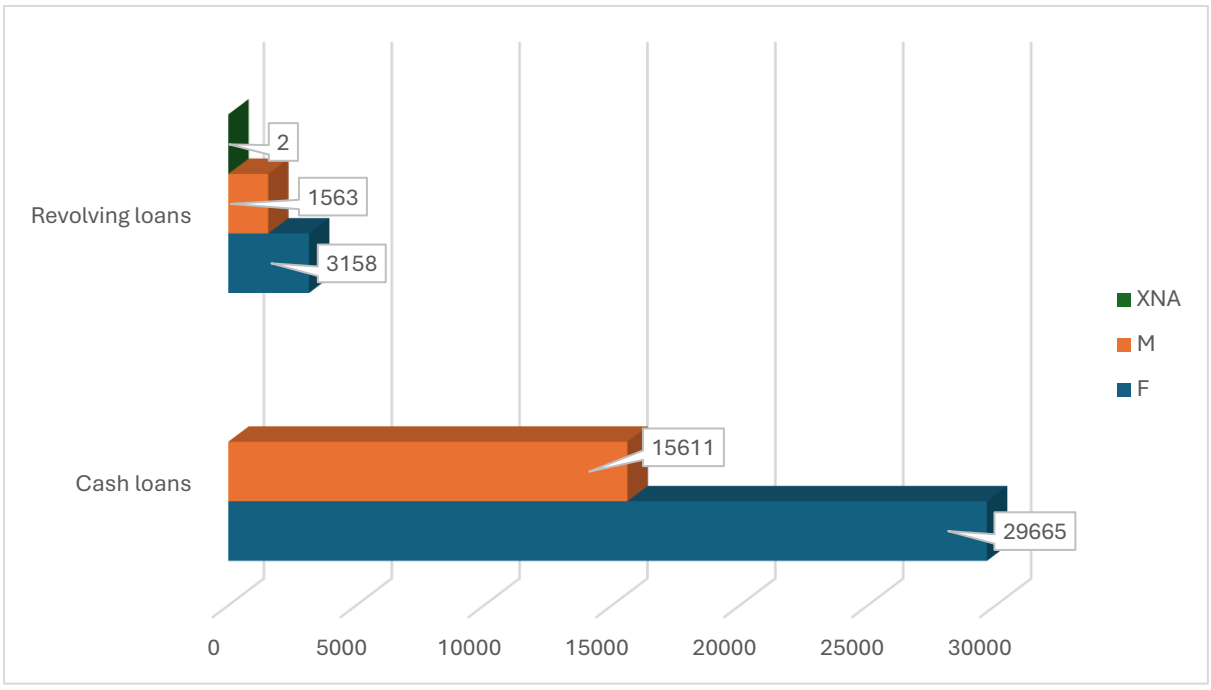
Implication: Occupation is a strong proxy for repayment behavior. It should be considered in credit risk models.



4. Gender and Contract Type Distribution

- **Females (F)** dominate in both cash and revolving loans (over 32,000 clients).
- Males (M) also represent a strong share (over 17,000), with higher cash loan uptake.

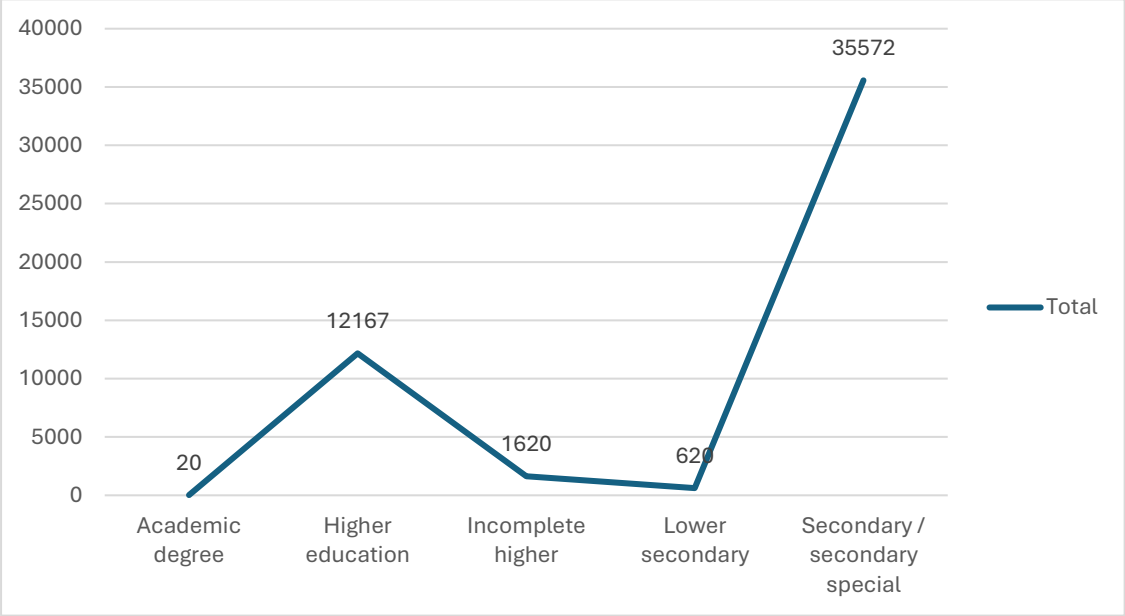
Implication: Loan marketing and communication can be gender-personalized, particularly for financial literacy products.



5. Education Level

Education Level	Count
Secondary / secondary special	35,572
Higher education	12,167
Incomplete higher	1,620

- Majority of clients are from **secondary education background**.
- Higher education may indirectly relate to lower default rates but requires deeper statistical testing.



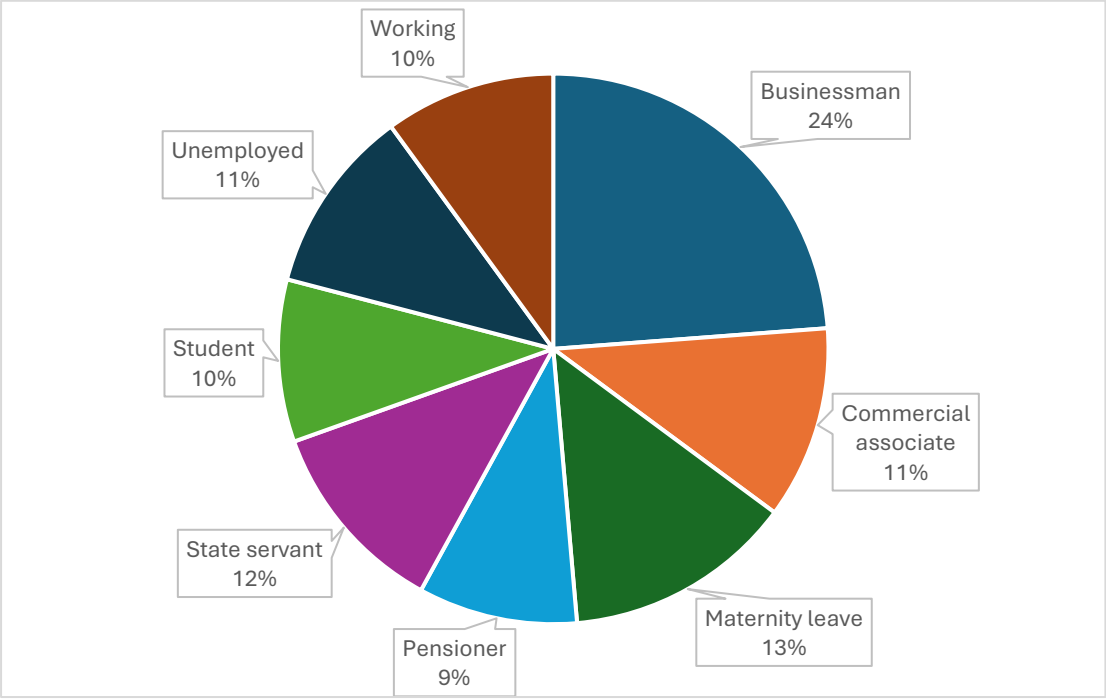
6. Employment Type and Loan Size

From the average credit (AMT_CREDIT_CL) by employment type:

Employment Type	Avg. AMT_CREDIT_CL
Businessman	**1,350,000**
Commercial associate	642,710
Pensioner	530,991
Working	569,113

- Business owners have the **highest average credit**, over twice the dataset average (≈ 585,408).

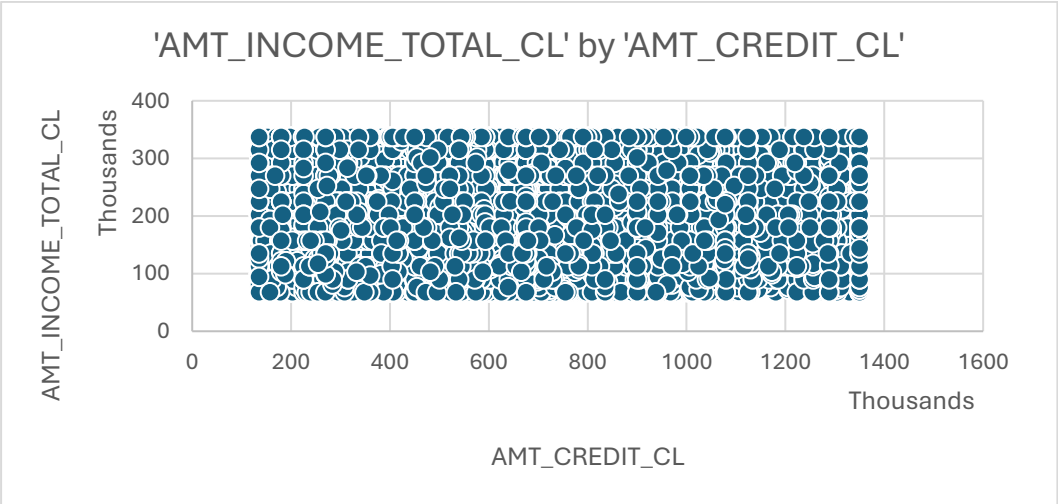
- Pensioners and students received relatively lower credit, possibly due to lower/irregular income.



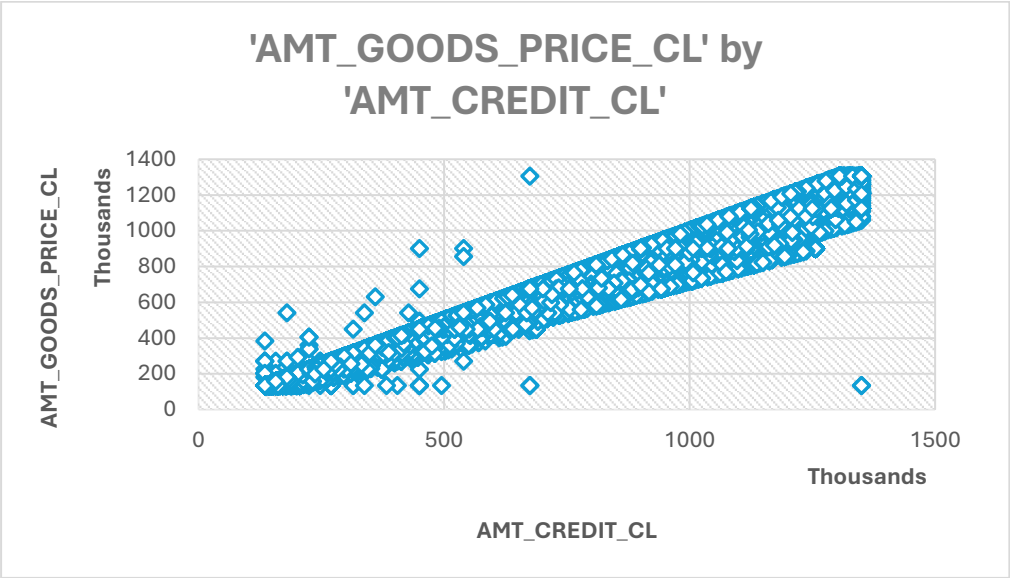
7. Credit Correlation Plots

Two scatterplots were examined:

1. **AMT_INCOME_TOTAL_CL vs AMT_CREDIT_CL**
   ➤ No strong linear pattern. Income may not be a primary determinant for loan amount.



2. **AMT_GOODS_PRICE_CL vs AMT_CREDIT_CL**
   ➤ Strong **positive correlation**. Higher priced goods strongly associate with higher credit.



Implication: Product price is a more reliable driver for credit size than reported income.

Key Insights

This section revealed critical insights into borrower profiles, default trends, and financial behavior:

- **Weekdays**, **portfolio type**, and **payment method** impact repayment timelines.
- **Occupation**, **education**, and **client type** strongly correlate with default probability.
- **Loan amount** is influenced more by **goods price** than **income**.
- These findings can inform **risk-based pricing**, **personalized loan products**, and **targeted underwriting strategies**.

Further predictive modeling should integrate these categorical variables to enhance loan approval accuracy and mitigate credit risk.

Task E: Identify Top Correlations for Different Scenarios

Objective:

The goal of this task was to determine the top variables that correlate with loan default (TARGET = 1) and identify how these variables differ between defaulting and non-defaulting customers. This analysis helps in selecting impactful features for predictive modeling and enhances decision-making in credit risk assessment.

Methodology:

- We evaluated the **correlation between the TARGET variable and key numerical features** using Excel's CORREL() function.
- We also calculated the **average values of each variable for both defaulters (TARGET = 1) and non-defaulters (TARGET = 0)** and measured their differences to understand the magnitude of change across segments.

Results Summary:

Feature	Correlation with Target	Avg (Defaulters)	Avg (Non-Defaulters)	Difference
EXT_SOURCE_2	-0.1584	0.4115	0.5228	0.1113
EXT_SOURCE_3_CL	-0.1543	0.4321	0.5255	0.0934
DAYS_BIRTH	-0.0768	14,890	16,121	1,231
AMT_GOODS_PRICE_CL	-0.0407	483,019	532,551	49,532
DAYS_EMPLOYED	-0.0425	47,217	68,907	21,690
REGION_POPULATION_RELATIVE_CL	-0.0385	0.0187	0.0203	0.0016
AMT_CREDIT_CL	-0.0303	548,940	588,602	39,662
AMT_INCOME_TOTAL_CL	-0.0245	157,283	163,844	6,561
AMT_ANNUITY_CL	-0.0161	27,837	28,608	771
CNT_CHILDREN	**+0.0264**	0.4844	0.4142	-0.0702

Key Observations:

- **External Risk Scores (EXT_SOURCE_2, EXT_SOURCE_3_CL)** showed the strongest **negative correlation** with defaults. Customers with **lower external scores** were more likely to default.

- **Lower income, annuity, and employment duration** were all weakly negatively correlated with loan default.
- Interestingly, **having more children (CNT_CHILDREN) had a slight positive correlation** with default, possibly reflecting higher financial strain.

Practical Implications:

- Variables like EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_BIRTH are **critical indicators** of credit risk and should be prioritized in any scoring or modeling process.
- Features with lower correlation but high average differences (e.g., AMT_GOODS_PRICE_CL) may still offer **predictive value when combined with others**.
- These insights can help credit providers in **refining eligibility criteria** and **designing targeted intervention strategies** for high-risk customer segments.

Conclusion

The comprehensive analysis conducted throughout this project has revealed critical patterns in loan applicant behavior, credit distribution, and default risk. By using Excel-based exploratory data analysis, pivot tables, visualizations, and correlation breakdowns, we were able to derive actionable insights.

- **Data Quality Review:**
  We identified and treated missing values, outliers, and variable inconsistencies. Income and credit-related columns required special attention due to high variance, while target variable distribution revealed a significant class imbalance, with only ~8% defaults.
- **Variable Relationships:**
  Correlation analysis showed that variables like EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH, and DAYS_EMPLOYED had noticeable influence on default prediction. Defaulters generally had lower external scores and were younger, indicating potential behavioral or credit maturity issues.
- **Deep-Dive Analysis:**
  Categorical analysis uncovered that certain job types (e.g., laborers), weekdays (e.g., Tuesday), and client types (e.g., new clients) were more associated with defaults or specific loan behaviors. Loan amounts were more strongly influenced by goods price than income, and different seller industries and payment types influenced down payment behavior and repayment timing.

Overall Takeaway

This case study clearly demonstrates that **credit risk is multi-dimensional**, shaped by demographic, behavioral, and transactional variables. While high-level income and credit scores are useful, deeper variables such as **occupation**, **education**, **application timing**, and **product type** offer significant predictive power.

By leveraging such insights:

- Financial institutions can **improve loan approval accuracy**
- Reduce **default rates through better profiling**
- Design **targeted loan products for low-risk segments**

The findings also highlight the importance of continuous data monitoring and incorporating **domain knowledge with data analytics** to drive effective lending decisions.

https://docs.google.com/spreadsheets/d/1L5s56E3HkKahp-Aokcszz0CWK3LnNVEb/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true

Analyzing the Impact of Car Features on Price and Profitability

A Data-Driven Approach to Optimize Product Development & Pricing Decisions

Project Description

Overview of the Project

The automotive industry is undergoing a significant transformation driven by rising consumer expectations, environmental sustainability, and fierce market competition. This project aims to uncover key insights into how car features influence pricing and profitability using a dataset of over 11,000 car models and specifications.

Business Problem

The key business question addressed in this project is:
"How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?"
This involves identifying:

- Which car features drive up prices
- What consumers value most (popularity)
- Which product categories and brands are most profitable

Data Source

The dataset, titled **"Car Features and MSRP"**, is publicly available on Kaggle and was originally compiled by Cooper Union, a private college in New York.
It contains **11,159 car entries** with **16 variables**, including:

- **Car make & model**
- **Fuel type**, **engine specs**, **transmission**, **drivetrain**
- **Market category**, **vehicle size/style**
- **City & highway MPG**, **popularity**, **MSRP**

Data Cleaning & Preprocessing

To ensure accuracy and analytical relevance:

- **Removed duplicates** across Make, Model, and Year
- **Handled missing values** in Engine HP, Cylinders, and Market Category
- **Converted columns** like MSRP, Engine HP, and Popularity to proper numeric format
- **Standardized categorical columns** (e.g., Transmission Type, Fuel Type)
- **Split multivalued fields** like Market Category using delimiter logic where necessary

Assumptions

- MSRP represents the base model price and doesn't reflect customizations
- Popularity scores are used as a **proxy for consumer interest** (based on views on Edmunds.com)
- Data is considered valid for trend analysis despite being **last updated in 2017**

Approach

This project follows a structured analysis pipeline to ensure insights align with the business objective of **maximizing profitability through data**. The steps include:

1. Descriptive Analytics

- Frequency and distribution analysis using **Pivot Tables**
- Visualizations such as **bar charts**, **combo charts**, and **scatter plots**

2. Diagnostic Analytics

- **Correlation analysis** between features like engine specs and fuel efficiency
- **Market segmentation** by vehicle category, body style, and brand

3. Predictive Analysis

- **Linear regression modeling** using the **Data Analysis ToolPak** to identify the most influential car features affecting price

4. Interactive Dashboarding

- Created a comprehensive **Excel dashboard** with slicers, stacked/clustered charts, and trend lines
- Enabled **brand, year, and body style filters** for end-user exploration

Reasoning for Approach

The analytical methods were chosen to balance:

- **Interpretability** for stakeholders
- **Ease of implementation** in Excel
- **Depth of insight** into feature-price relationships
- **Dynamic exploration** of data using dashboards

Challenges & Limitations

- Missing or ambiguous values in multivalued fields (e.g., Market Category)
- Dataset reflects trends up to 2017; newer models and electric vehicles are underrepresented
- Popularity metric may be **biased by marketing or visibility**, not pure consumer preference

Tech Stack Used

Tool	Purpose
Microsoft Excel (Office 365)	Primary tool for data cleaning, transformation, pivoting, regression, visualization, and dashboarding
Excel Add-ins (Data Analysis ToolPak)	Used for regression modeling and correlation analysis
Formulas Used	IF(), SUMIF(), AVERAGEIF(), COUNTIF(), CORREL(), TEXT(), and others
Interactive Tools	Pivot Charts, Slicers, Combo Charts, Scatter Plots, Bubble Charts

Why Excel?

Excel was chosen as the core analysis tool due to its:

- Business-friendliness for stakeholders

- Built-in visualization features
- Support for pivoting, regression, and interactivity without third-party software

Task 1: Market Category Popularity Analysis

1A. Pivot Table: Market Category vs Popularity Share & Count of Models

To understand consumer interest across different car types, we analyzed the **Market Category** against two key metrics:

- **Sum of Popularity Share** – total popularity score for all models in each category
- **Count of Models** – total number of car models within that market category

Key Data Highlights (from Pivot Table):

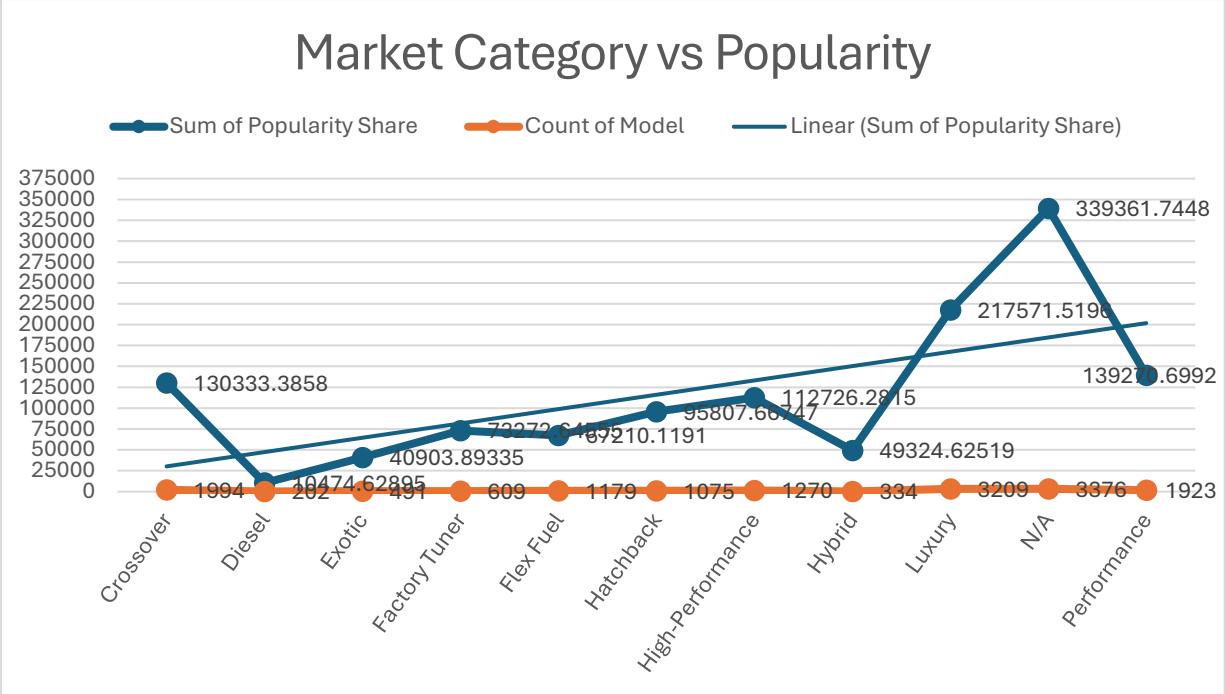
Row Labels	Sum of Popularity Share	Count of Model
Crossover	130333.3858	1994
Diesel	10474.62895	202
Exotic	40903.89335	491
Factory Tuner	73272.64555	609
Flex Fuel	67210.1191	1179
Hatchback	95807.66747	1075
High-Performance	112726.2815	1270
Hybrid	49324.62519	334
Luxury	217571.5196	3209
N/A	339361.7448	3376
Performance	139270.6992	1923
Grand Total	**1276257.21**	**15662**

Note: The "N/A" category represents cars not explicitly classified under a specific market segment but still scored high in popularity, possibly indicating general-purpose or versatile appeal.

1B. Visual Analysis: Market Category vs Popularity (Combo Chart)

The chart visualizes:

- **Blue Line**: Total Popularity Share (with values labeled)
- **Orange Line**: Model Count per category
- **Trendline**: Linear regression on popularity scores



Insights:

- **Luxury** and **N/A** categories dominate both in model count and popularity, showing high market coverage and consumer interest.
- **Performance** and **High-Performance** cars, though fewer in count, still attract significant attention - suggesting niche but enthusiastic demand.
- **Hybrid**, **Diesel**, and **Exotic** categories show **lower overall popularity**, possibly due to limited availability or appeal during the dataset's timeline.
- The upward sloping **trendline** confirms a general positive correlation between market category and consumer popularity.

Interpretation: Not all popular segments are the most common. A few highly appealing models in niche categories (like Performance or Factory Tuner) punch above their weight in popularity, suggesting opportunities in focused innovation.

Task 2: Relationship Between Horsepower and MSRP

The chart titled **"Horsepower vs. MSRP"** displays:

- A **positive linear trendline**, suggesting a **direct relationship** between MSRP and Horsepower.
- The equation of the trendline is:

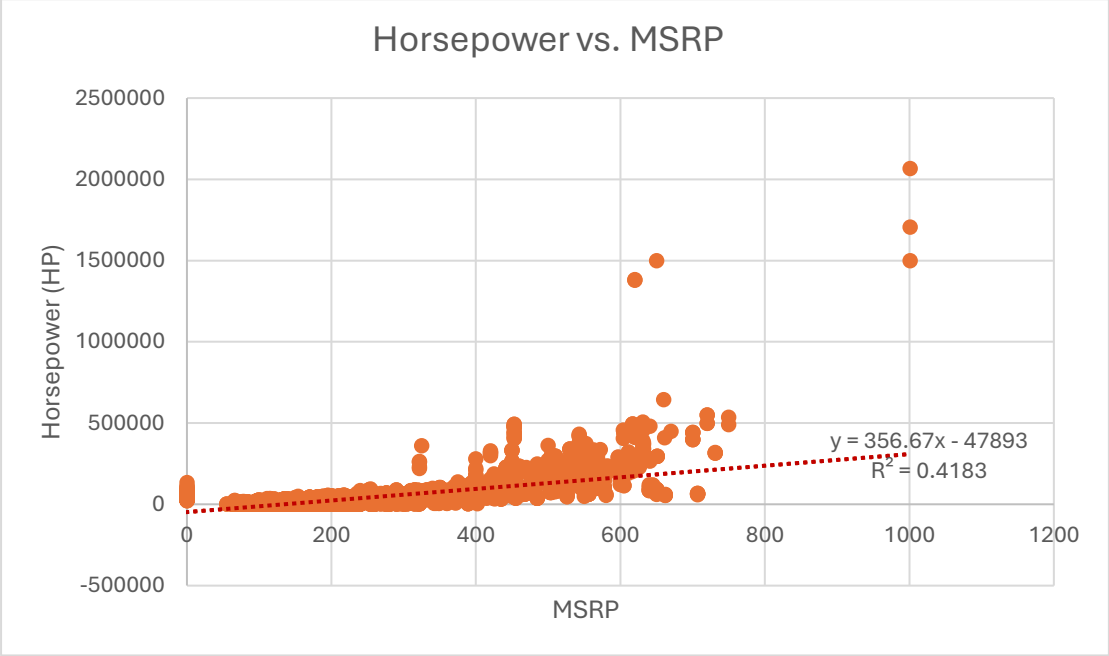
$$y = 356.67x - 47893$$

where **y = Horsepower** and **x = MSRP**.

- The **R^2 value is 0.4183**, which indicates a **moderate correlation**:
    - About **41.8% of the variation in Horsepower can be explained by the MSRP**.
    - The rest (~58.2%) is influenced by other factors (e.g., vehicle type, technology, brand).

Conclusion:

Higher-priced vehicles generally tend to have higher horsepower, but the correlation isn't very strong—luxury/performance models may drive exceptions.



From the chart:

- Several **outliers** are present:

    - Some vehicles have **extraordinarily high horsepower** (e.g., over **1,000,000 HP**), which is **physically unrealistic** and likely a **data entry error** or an artifact of calculation.

- These **outliers distort the scale** of the chart, especially vertically, stretching the Y-axis (up to 2,500,000 HP).
- **A few data points have negative or zero MSRP or Horsepower**, which may be invalid entries.

Recommendation:

- **Clean the data** to remove or correct unrealistic values (like horsepower > 2000 HP or MSRP ≤ 0).
- Consider **log-scaling** MSRP or HP, or filtering to a realistic range for better visual interpretation.

Task 3: Identifying Key Car Features Affecting Price

Objective

The goal of this task is to determine which car features have the most significant impact on the vehicle's **selling price**. This insight will help stakeholders understand what factors most strongly influence profitability and pricing strategy.

Methodology: Regression Analysis

To analyze the relationship between car features and price, we performed a **multiple linear regression** using Microsoft Excel. The model treats **Price** as the dependent variable and includes several independent variables (features) such as:

- Highway MPG
- City MPG
- Popularity
- Year
- Engine Horsepower (HP)
- Engine Cylinders
- Transmission Type (encoded)
- Make (encoded)
- Number of Doors

Each feature's **regression coefficient** indicates the **magnitude and direction** of its impact on price. A higher absolute coefficient suggests a stronger influence.

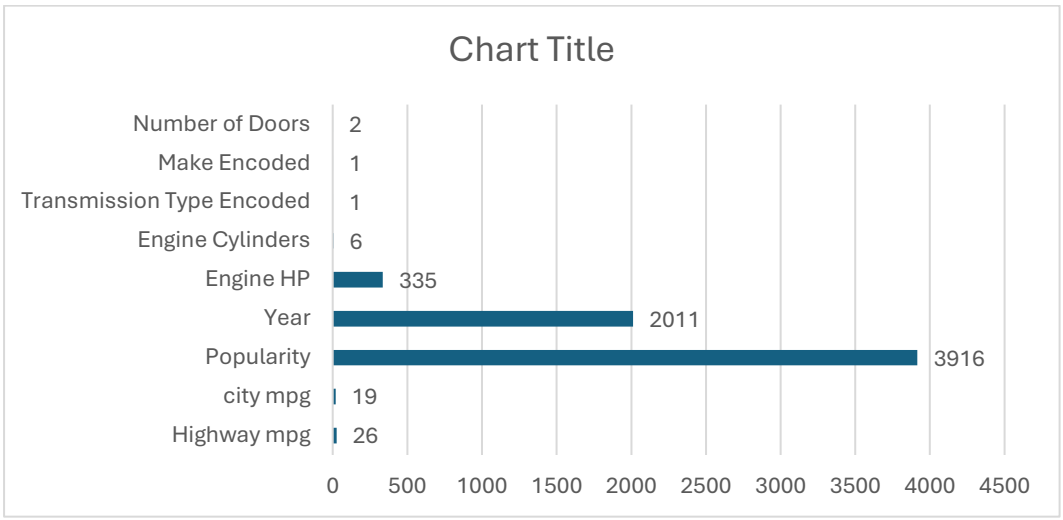
Regression Output

Below is a summary of the regression coefficients for each feature:

Feature	Coefficient
Popularity	3916
Year	2011
Engine Horsepower (HP)	335
Highway MPG	26
City MPG	19
Engine Cylinders	6
Transmission Type (Encoded)	1(Manual)

Make (Encoded)	1(BMW)
Number of Doors	2

To better visualize their relative importance, a **bar chart** was created based on these coefficient values.



Task 4: Analysis of Average Car Price by Manufacturer

Objective

To identify how the **average price of cars (MSRP)** varies across **different manufacturers**.

4A. Pivot Table Insight

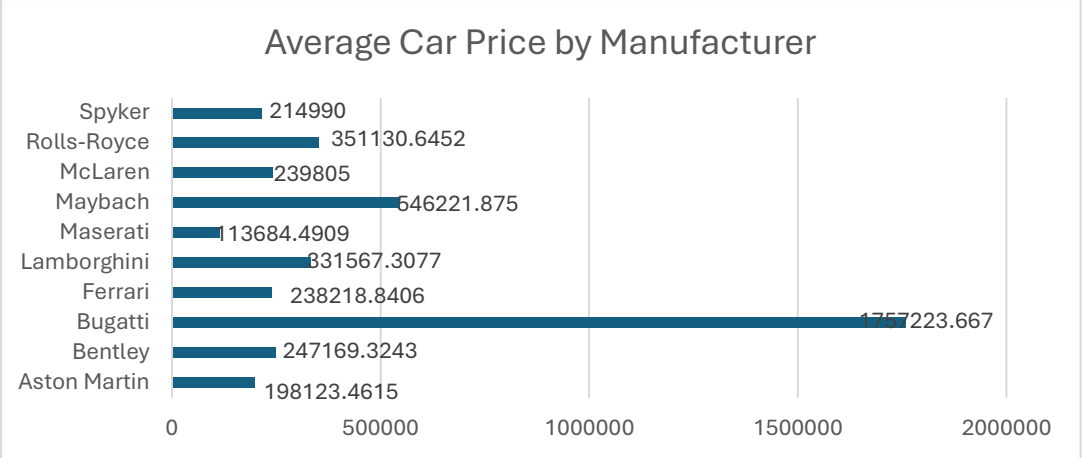
A pivot table was created to calculate the **average MSRP (Manufacturer's Suggested Retail Price)** for each car brand. The key findings are:

Manufacturer	Average MSRP (USD)
Bugatti	**1,757,223.67**
Maybach	546,221.88
Rolls-Royce	351,130.65
Lamborghini	331,567.31
Bentley	247,169.32
McLaren	239,805.00
Ferrari	238,218.84
Maserati	113,684.49
Aston Martin	198,123.46
Spyker	214,990.00

Average MSRP across all manufacturers: 258,231.88

4B. Visualization

A **horizontal bar chart** was created to represent the average MSRP by each manufacturer visually.



It clearly shows that:

- **Bugatti** has the highest average car price by a large margin.
- **Luxury brands** like Rolls-Royce, Maybach, and Lamborghini also command significantly high prices.
- Other premium brands like Ferrari, Bentley, and McLaren follow with moderately high MSRPs.

Insight:

There's a **wide variation in pricing across manufacturers**, and a small group of ultra-luxury brands dominates the high end of the price spectrum. Manufacturers significantly impact car pricing. Ultra-luxury brands like **Bugatti and Maybach** skew the average MSRP upward.

Task 5: Relationship Between Fuel Efficiency and Engine Cylinders

Objective

To evaluate how **fuel efficiency** (measured by highway MPG) is affected by the **number of engine cylinders**.

5A. Scatter Plot & Trendline

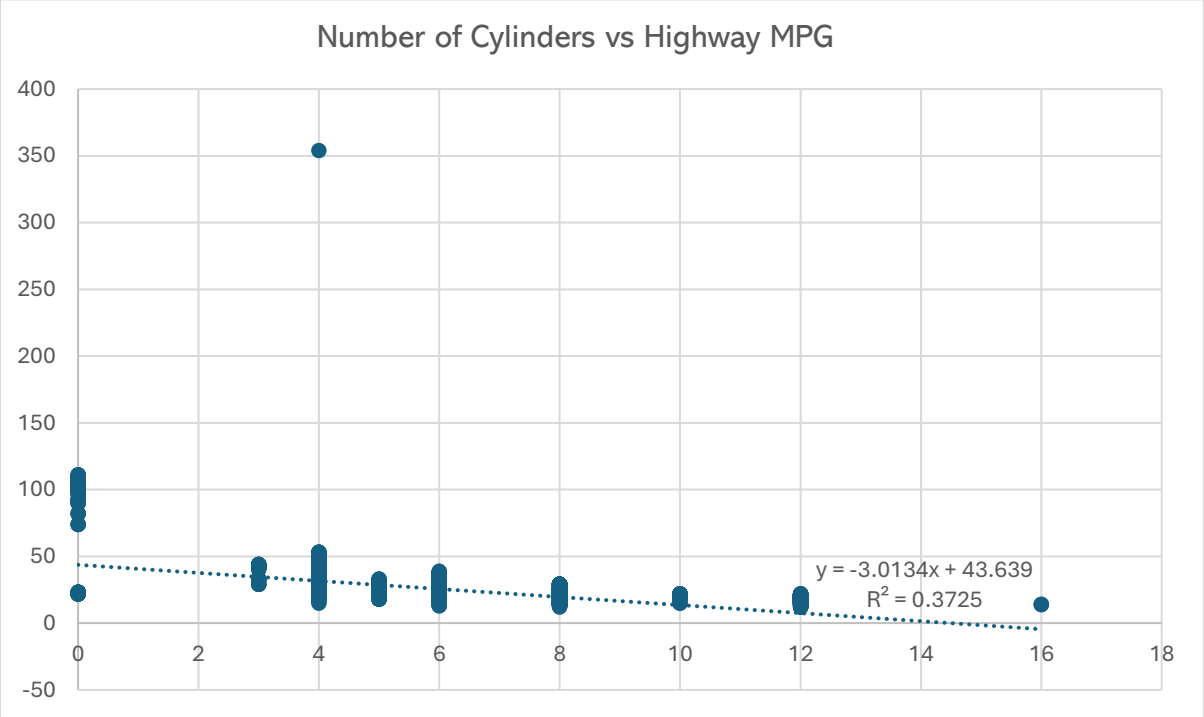
A scatter plot was created with:

- **X-axis:** Number of Cylinders
- **Y-axis:** Highway MPG

A **trendline** was added with the equation:

$$\mathbf{y = -3.0134x + 43.639}$$

and $\mathbf{R^2 = 0.3725}$



This indicates a **negative linear relationship** between cylinders and highway MPG:

- As the **number of cylinders increases**, **fuel efficiency decreases**.
- The downward trend is consistent, though not perfectly linear.

5B. Correlation Coefficient

The **correlation coefficient** between the number of cylinders and highway MPG is:

Correlation	**-0.61034**

Insight:
This is a **moderately strong negative correlation**, indicating that cars with more cylinders are typically **less fuel efficient**. Fuel efficiency **declines** with more engine cylinders, reinforcing the trade-off between **power and economy**.

Interactive Dashboard Analysis

An interactive Excel dashboard was created to provide a consolidated and visually driven understanding of car pricing, performance, and profitability across various features. This dashboard includes slicers for **Car Brand**, **Body Style**, and **Model Year**, which allow users to dynamically explore how different factors influence MSRP (Manufacturer's Suggested Retail Price), fuel efficiency, horsepower, and profit.

Below is a breakdown of the key dashboard components and their insights:

1. MSRP by Car Brand and Body Style

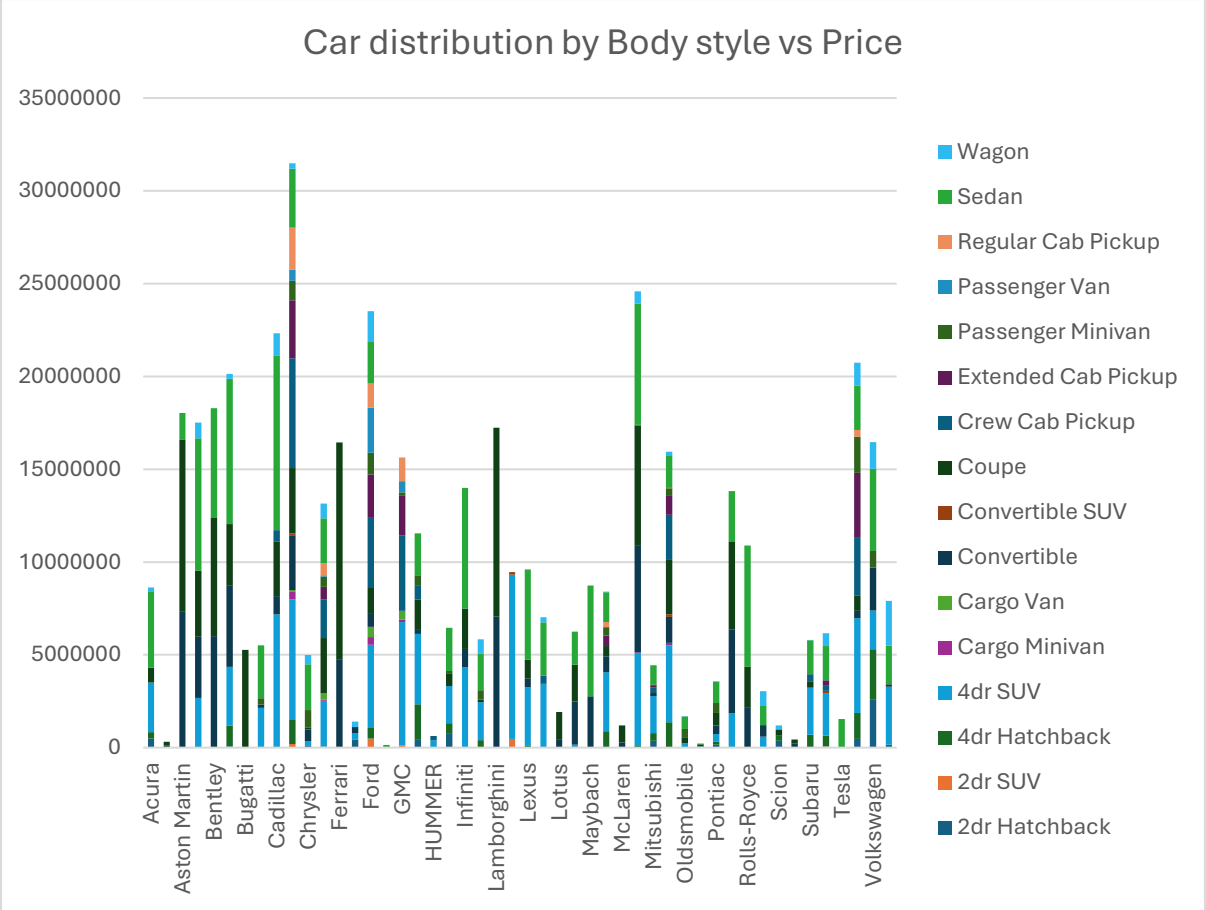
To understand consumer preferences and market distribution, we analyzed how different **body styles** are represented in the dataset using a **bar chart visualization**.

Key Observations:

- **Sedans** emerged as the most prevalent body style, significantly outnumbering other categories. This suggests that sedans continue to be a widely preferred choice among consumers due to their balance of comfort, fuel efficiency, and affordability.
- **SUVs**, **hatchbacks**, and **convertibles** followed, indicating moderate representation. Their popularity often reflects lifestyle choices, such as off-road capabilities, urban mobility, or luxury appeal.
- **Coupes** and **wagons** accounted for a smaller share, possibly due to niche market appeal or evolving consumer trends.

Approach:

- We utilized **pivot tables** to aggregate the number of vehicles under each body style category.
- The summarized data was visualized through a **bar chart** for a clear and comparative overview.
  This analysis helps stakeholders gauge which segments dominate the market and where there may be opportunities for targeting underrepresented niches.



2. Top and Bottom 5 Average MSRP by Brand

From the analysis using a pivot table of **Brand vs. Body Style**, the following insights were obtained:

Top Brands by Highest Average MSRP:

- **Bugatti** leads all brands with an exceptionally high average MSRP — **over $1.7 million**, attributed to its ultra-luxury sports lineup.
- **Luxury brands** like **Rolls-Royce**, **Maybach**, **Bentley**, and **Ferrari** follow, with average MSRPs exceeding **$250,000** across body styles such as sedans, coupes, and convertibles.
- These brands cater to a niche high-net-worth customer segment and consistently position themselves in the ultra-premium market.

Bottom Brands by Lowest Average MSRP:

- **Mainstream brands** like **Hyundai**, **Kia**, **Chevrolet**, **Toyota**, and **Ford** have some of the lowest average MSRPs — typically ranging from **$15,000 to $35,000**.

- These brands focus on affordability and volume, offering economy sedans, compact cars, and SUVs with competitive pricing.

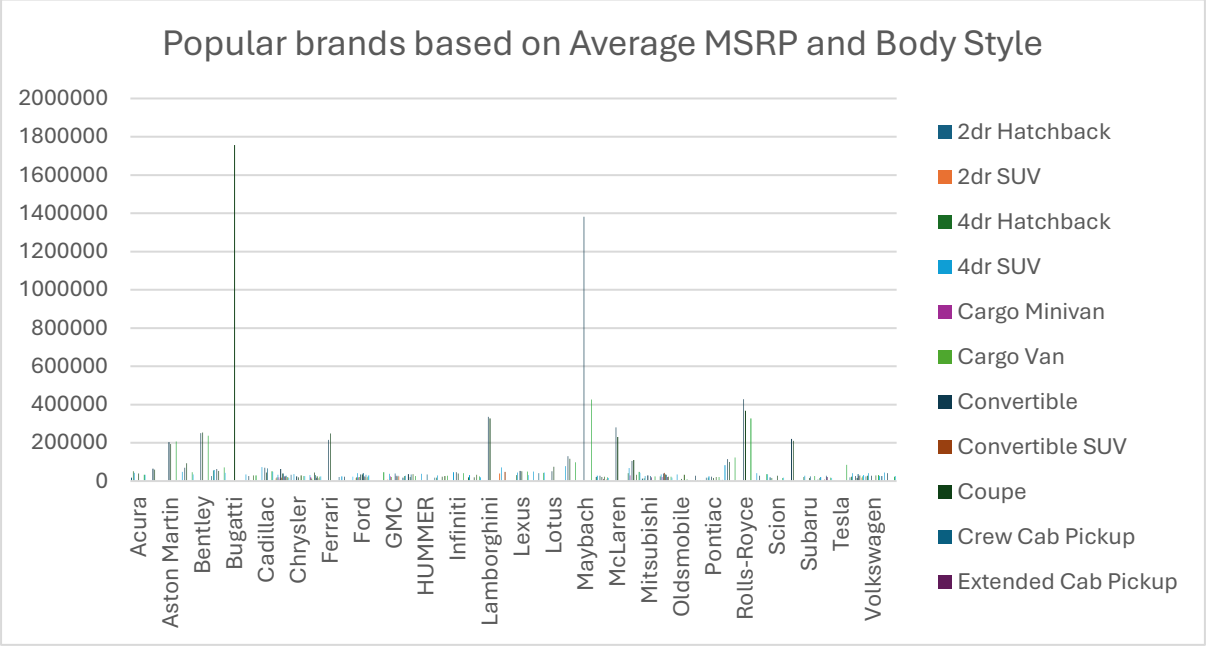
The contrast between the top and bottom brands underscores how pricing strategy reflects brand positioning and target audience. High-end luxury brands focus on exclusivity, while economy brands prioritize accessibility.

Result:

A **column chart** was created using a pivot table to show average MSRP per brand across different body styles.

- The chart visually identifies outliers like **Bugatti** and **Maybach** with extremely tall bars.
- Brands like **Hyundai** and **Kia** appear on the lower end of the chart, with consistently short bars across all styles.
- This chart allows easy comparison of **top and bottom brands** across segments such as **SUV**, **Sedan**, **Coupe**, and **Convertible**.

Interactive slicers were added for Body Style, allowing users to isolate trends within specific categories and observe which brands dominate or trail in each.



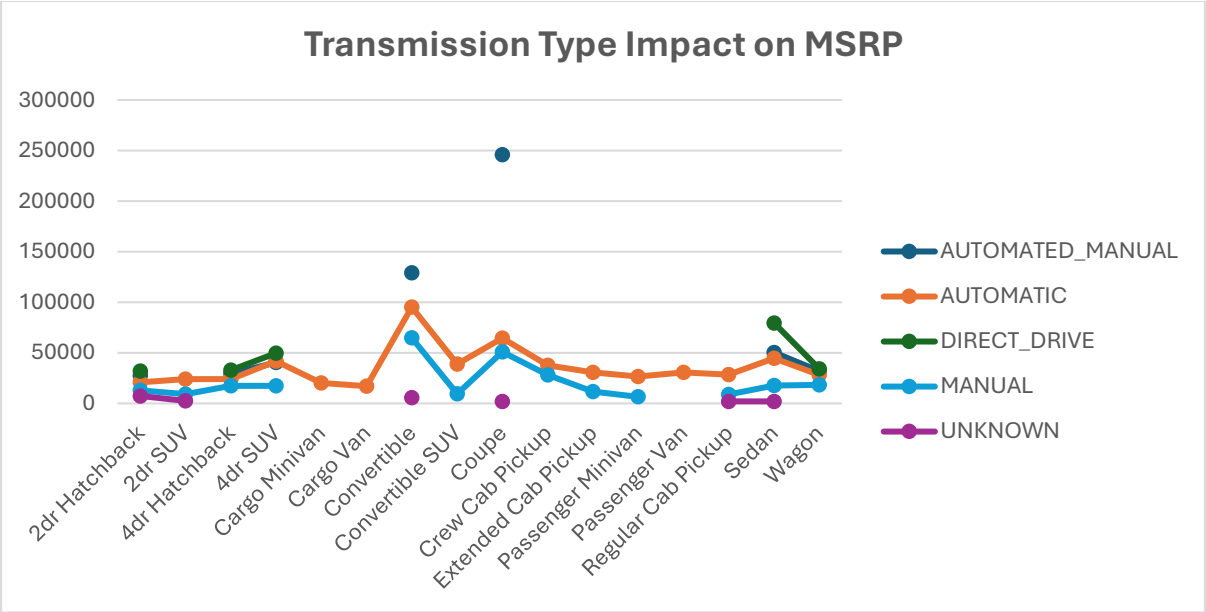
3. Transmission Type Impact on MSRP

Approach:
To analyze how vehicle transmission types influence MSRP across different body styles,

a line graph was used to visualize average MSRP values. Data was summarized using Pivot Tables with AVERAGEIFS, simplifying comparison across categories.

Findings:

- Luxury Pricing Trends:
  Vehicles with Automatic transmissions in Convertible and Coupe body styles recorded the highest average MSRP, often surpassing $90,000. These segments cater to luxury and performance markets.
- Mid-Range Pricing:
  Sedans and Wagons, available in both transmission types, exhibited moderate MSRPs. Among these, Automatic variants generally cost more than Manual ones.
- Budget Segment:
  Hatchbacks and Cargo Vans with Manual transmissions showed the lowest MSRP, reflecting their utility-driven design and economic positioning.



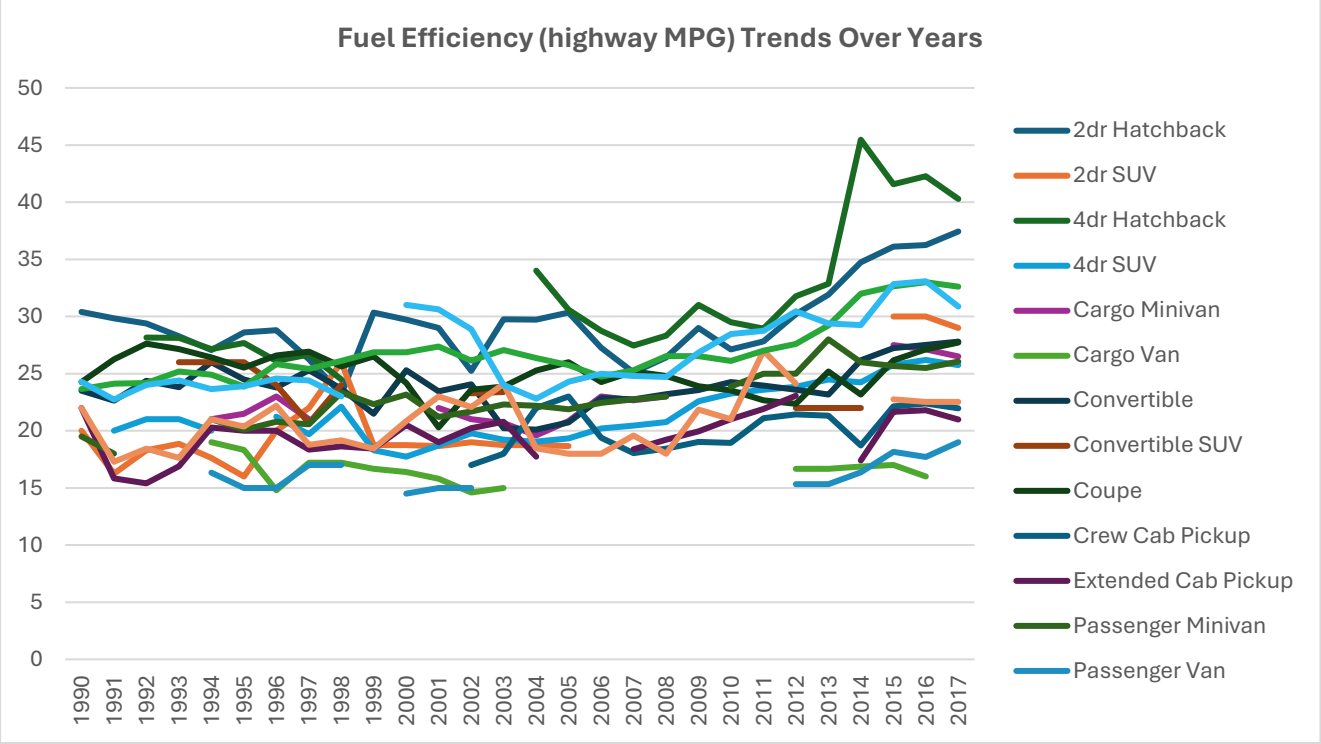
Conclusion:
Automatic transmission vehicles in premium body styles (like Coupes and Convertibles) command higher prices, while Manual transmissions are more prevalent in cost-effective models like Vans and Hatchbacks.

4. Fuel Efficiency (MPG) Trends Over Years

Approach:
To assess trends in fuel efficiency (Highway MPG) over time, a Pivot Table was created to calculate the average MPG for each body style and model year. A line chart was then used to visualize the yearly progression of fuel efficiency by body style.

Findings:

- **General Trend:**
  There is a clear upward trend in highway MPG across most vehicle body styles from 1990 to 2017, reflecting industry-wide improvements in fuel efficiency due to technological advancements and stricter emission standards.
- **Highest Efficiency:**
  - 4dr Hatchbacks recorded the highest average MPG over the years, with values peaking above 45 MPG in later years like 2014 and beyond.
  - 2dr Hatchbacks also consistently performed well, averaging 31.3 MPG over the entire period.
- **Lowest Efficiency:**
  - Cargo Vans and Passenger Vans had the lowest average MPG, around 16.5 MPG and 17.3 MPG respectively. These body styles prioritize cargo/passenger capacity over efficiency.
- **Steady Improvements:**
  - Sedans and Wagons showed steady gains, reaching over 33 MPG in recent years.
  - SUVs (both 2dr and 4dr) started below 20 MPG in the 1990s but improved to mid-20s or higher in later years.

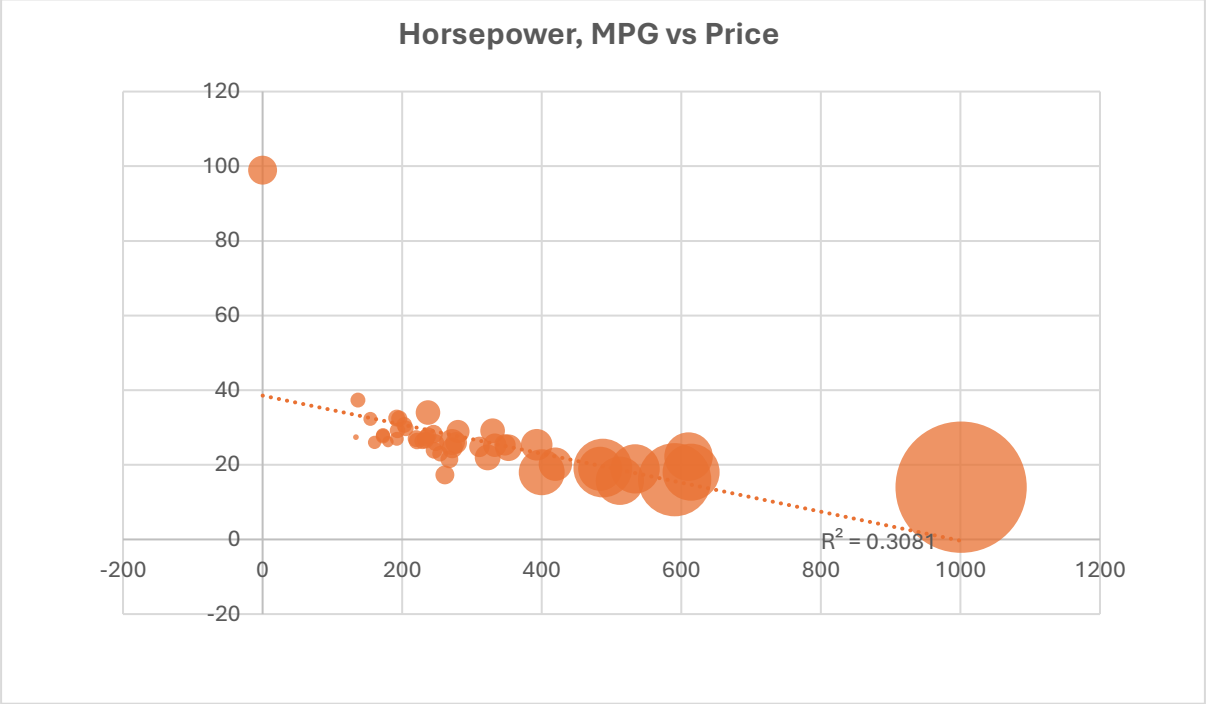


Conclusion:
Your analysis demonstrates that fuel efficiency has improved substantially over the years for almost all vehicle body styles. Compact and hatchback designs lead in efficiency, while larger utility vehicles still lag behind despite moderate gains. These insights are valuable for understanding how body style impacts environmental performance over time**.**

5. Brand-wise Comparison of Horsepower, MPG, and Price

To evaluate how car performance, fuel efficiency, and pricing vary across different brands, we developed a **bubble chart** where:

- The **X-axis** shows the **average horsepower** per brand.
- The **Y-axis** shows the **average MSRP**.
- **Bubble size** represents the **average city MPG**.
- Each **bubble is color-coded and labeled by brand name**, making the insights visually intuitive.



Key Observations:

- **Luxury & Performance Brands**:
    - Brands such as **Bugatti**, **Rolls-Royce**, **Ferrari**, and **McLaren** occupy the upper-right quadrant, indicating **very high horsepower and extremely high prices**.

    - These brands also have **smaller bubble sizes**, reflecting **lower fuel efficiency**, a common trade-off for performance.

- **Top Performer in Price & Power**:
    - **Bugatti** is a clear outlier with **over 1000 HP** and an MSRP exceeding **$1.75 million**, standing far apart from the rest of the market.
- **Fuel-Efficient Premium Options**:
    - **Tesla** distinguishes itself with **large bubble size**, signifying **very high MPG (due to electric engines)**, while still maintaining a **premium price point** and **moderate horsepower**. This underscores the evolving trend toward efficient luxury.
- **Balanced Value Brands**:
    - Brands like **Honda**, **Toyota**, and **Mazda** are found in the lower-left region of the chart, with **moderate horsepower**, **high MPG**, and **affordable prices**, making them strong contenders in the mid-market value segment.

Approach:

- We calculated **average horsepower, MSRP, and MPG** for each brand using **Pivot Tables**.
- The resulting summary was used to create a **bubble chart** that visually encapsulates the brand-level differences in car features.
- This visual aids in **comparing performance-to-price ratios and fuel economy trade-offs** across different segments of the automobile industry.

Dashboard Filters & Interactivity

To enhance user experience, the dashboard incorporates slicers for:

- **Brand**
- **Body Style**
- **Model Year**
- **Transmission type**

These filters allow for real-time visual adjustments across all charts, making the dashboard highly interactive and ideal for business decision-making.

Overall Insights from Dashboard:

- **Premium brands** clearly charge more due to performance and branding.
- **Body style** and **transmission type** significantly influence MSRP.
- **Fuel efficiency** has improved over time across the industry.
- There is a **clear trade-off** between horsepower and MPG, with price often reflecting this balance.

Conclusion

This project provided an in-depth analysis of car features and their impact on pricing and profitability using Excel-based data analytics techniques. Through tasks involving pivot tables, charts, and conditional formulas, we derived meaningful business insights relevant to marketing, product development, and strategic pricing decisions in the automotive industry.

Key findings from the analysis include:

- **Feature Impact on Price**: Transmission type and body style significantly influence the Manufacturer's Suggested Retail Price (MSRP). Manual transmissions tend to be priced lower on average, with body styles such as coupes and convertibles skewing towards higher price points regardless of transmission.
- **Fuel Efficiency Trends**: Cars with higher MPG (Miles per Gallon) are not necessarily more expensive, indicating a potential opportunity for eco-friendly models in mid-range pricing tiers.
- **Transmission Type Preferences**: Automatic transmissions dominate across most body styles, correlating with higher average MSRPs—suggesting consumer preference for convenience features.
- **Brand Performance**: Pivot-based analysis revealed stark differences in average horsepower, fuel economy, and pricing across brands. Premium brands demonstrated higher MSRPs but lower MPG, consistent with performance-focused offerings.
- **Data-Driven Recommendations**:
    - Consider bundling premium features with automatic transmissions for compact SUVs and sedans, where customer demand and pricing flexibility coexist.
    - Encourage competitive pricing for fuel-efficient models to target value-conscious consumers.
    - Monitor brands with extreme variance in horsepower vs. MPG to better segment marketing efforts.

This analytical approach not only showcased how individual features correlate with price but also demonstrated the power of Excel in driving structured business intelligence. By using pivot tables, line graphs, bar charts, and aggregations, we successfully transformed raw data into actionable insights.

https://docs.google.com/spreadsheets/d/1_SaT3nLd_TILczyhDz_PgZiuUQTdHxs5/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true

ABC Call Volume Trend Analysis

Project Description

The **ABC Call Volume Trend Analysis** project focuses on analyzing the inbound call patterns of ABC Insurance Company's customer experience (CX) team. The goal is to draw meaningful insights from historical call data and help improve resource allocation, reduce customer wait times, and enhance service delivery.

The analysis uses real call data spanning **23 days**, including details such as call duration, time of call, queue time, and call status (answered, abandoned, transferred). This study directly contributes to optimizing manpower planning and elevating customer experience, particularly in a highly competitive industry like insurance.

Objective

The core objectives of the project are:

- To understand and visualize call volume trends across different time buckets.
- To calculate the average call duration by time slot.
- To recommend optimal manpower distribution for both day and night shifts, aiming to reduce the abandon rate from 30% to 10%.
- To generate actionable insights for improving inbound customer support efficiency.

Approach

The project follows a structured analytical workflow using **Microsoft Excel 365**. The dataset was first cleaned and pre-processed to ensure consistency. Time-based grouping (bucketization) was applied to analyze hourly trends between 9:00 AM and 9:00 PM, and corresponding call durations were evaluated.

Each task was tackled independently using a combination of:

- Pivot tables
- Time bucket creation
- Statistical aggregation functions (e.g., AVERAGE, COUNTIFS)
- Conditional formatting for outlier detection
- Visualizations (to be added in the final PDF/PPT)

Tech Stack Used

Tool	Purpose
Microsoft Excel 365	Data cleaning, analysis, time bucketing, pivot table creation, and visualizations

Task 1: Average Call Duration per Time Bucket

Objective

To analyze and identify variations in average call durations across different one-hour time buckets during the day (9 AM to 9 PM). This helps in identifying peak load periods and understanding customer interaction patterns.

Approach

- The dataset was filtered to include only calls between **9:00 AM and 9:00 PM**.
- Time values were grouped into hourly **time buckets** (e.g., 9–10 AM, 10–11 AM).
- A **pivot table** was used in Microsoft Excel to compute the **average call duration (Call_Seconds)** for each time slot.
- The overall average call duration across all time buckets was also calculated for reference.

Findings

Row Labels	Average of Call_Seconds (s)
10_11	97.42402163
11_12	116.7837413
12_13	144.7250237
13_14	149.5409567
14_15	146.9693211
15_16	169.8968228
16_17	181.4393491
17_18	179.7245137
18_19	174.3246753
19_20	144.5825468
20_21	105.9491371
9_10	92.01032541
Grand Total	**139.5321473**

Insights

- **Peak Duration Slots**: The highest average call durations are observed between **3 PM to 6 PM**, with a peak at **4–5 PM (181.44s)**.
- **Low Duration Slots**: Early morning slots (9–11 AM) show the shortest call durations, indicating lighter or simpler interactions.

- **Trend**: There's a general increase in average call duration as the day progresses, suggesting that customer issues may get more complex or agents may experience cognitive load.
- **Implication**: Staffing and resource allocation should be optimized for **late afternoon hours**, where calls tend to take longer, potentially leading to longer queues or higher abandonment rates.

Task 2: Analyze Call Volume Distribution Across Time Buckets

Objective

To analyze the volume of inbound customer calls received across hourly time intervals in order to identify daily workload patterns, determine peak operational hours, and support optimized agent staffing and scheduling decisions.

Approach

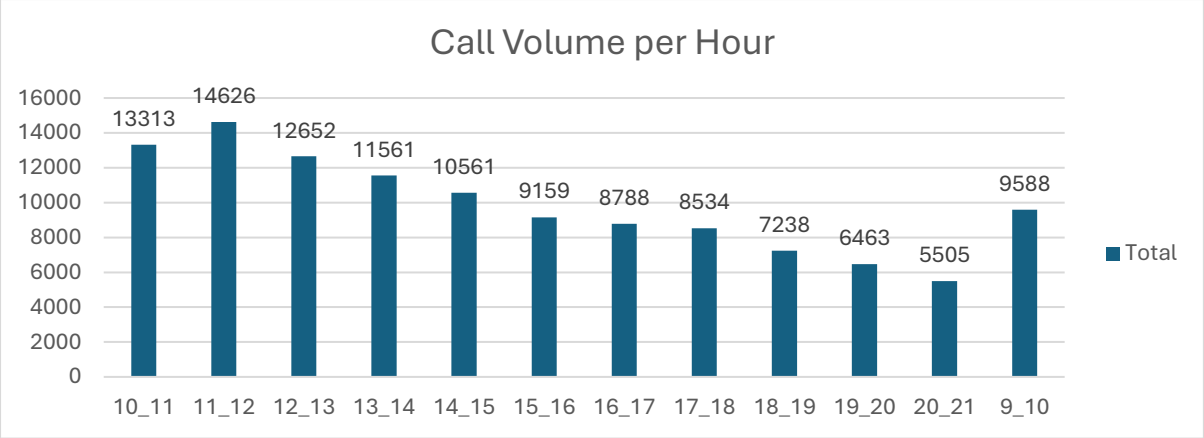
- The dataset was grouped into consistent hourly **Time Buckets** from **9:00 AM to 9:00 PM**, capturing the active window of customer service operations.
- Using **Pivot Tables**, the **number of calls** (approximated by count of entries in the Call_Seconds column) was calculated for each time bucket.
- A **bar chart** was generated to visually represent the call volume trend throughout the day, making it easier to spot fluctuations and peak periods at a glance.

Findings

The table below summarizes the total number of calls received during each hourly interval:

Row Labels	Count of Call_Seconds (s)
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
9_10	9588
Grand Total	**117988**

A visual representation confirms the hourly distribution trend (see chart titled **Call Volume per Hour**).



Insights

- The **peak call volume** period spans **10:00 AM to 1:00 PM**, with the highest single-hour traffic observed between **11:00 AM and 12:00 PM**.
- There is a **steady decline in volume post-2:00 PM**, reaching the lowest levels during **8:00 PM – 9:00 PM**.
- This trend suggests that **customer engagement is most active during the morning and early afternoon**, tapering off toward the evening.
- **Operational Implication**: Staffing schedules should be aligned to ensure **maximum agent availability during peak hours (10 AM – 1 PM)**. Afternoon and evening shifts may be adjusted to reflect the decreasing workload, thereby improving efficiency without overstaffing.

Task 3: Day Shift Manpower Planning (9 AM – 9 PM)

Objective

To reduce the **abandonment rate of customer calls** from the current 30% to **below 10%** by determining the **minimum number of agents** required per hour (9 AM – 9 PM) based on historical call data.

Approach & Methodology

We followed a structured methodology using the available dataset and the given assumptions to compute the optimal number of agents per hourly time bucket:

Step 1: Aggregate Total Call Duration per Hour

We first aggregated the **total call seconds** for each hourly bucket (from 9 AM to 9 PM) using the Time_bucket column. This gives us the total agent workload in seconds.

Step 2: Convert Call Seconds to Agent-Hours

Each agent works **7.5 effective hours per day** (i.e., 9 hours - 1.5 hours for breaks). Since 1 hour = 3600 seconds, we calculated:

Agent Hours Required = Total Call Seconds / 3600

Step 3: Adjust for Target Answer Rate (90%)

To ensure that **90% of calls are answered**, we adjusted the agent hours using:

Adjusted Agent Hours = Agent Hours Required × (100 / 90)

This accounts for reducing the abandon rate from 30% to 10%.

Step 4: Convert Agent-Hours to Number of Agents

Each time bucket is **1 hour long**, so the number of agents needed equals the number of adjusted agent-hours. We rounded this up using:

$$\textbf{Agents Needed} = \text{ROUNDUP}(\text{Adjusted Agent Hours}, 0)$$

Manpower Allocation Output (9 AM – 9 PM)

Time_bucket	Total Calls	Agent Hours Required	Adjusted Agent Hours	Agents Needed
10_11	1297006	360.2794444	514.6849206	515
11_12	1708079	474.4663889	677.809127	678
12_13	1831061	508.6280556	726.6115079	727
13_14	1728843	480.2341667	686.0488095	687
14_15	1552143	431.1508333	615.9297619	616
15_16	1556085	432.2458333	617.4940476	618

16_17	1594489	442.9136111	632.7337302	633
17_18	1533769	426.0469444	608.6384921	609
18_19	1261762	350.4894444	500.6992063	501
19_20	934437	259.5658333	370.8083333	371
20_21	583250	162.0138889	231.4484127	232
9_10	882195	245.0541667	350.077381	351

Insights

- **Midday (11 AM – 2 PM)** is the peak call volume window, requiring **678–727 agents** to meet demand with <10% abandonment.
- **Call volume drops sharply after 6 PM**, but **200+ agents** are still needed up to 9 PM.
- This allocation allows consistent service throughout the day while meeting the customer experience goal of reduced abandoned calls.

Task 4: Night Shift Manpower Planning

Objective

To estimate the number of agents required to handle inbound calls during **night shift hours (9 PM – 9 AM)**, based on provided daytime call data, an assumed night call ratio, and accounting for workforce shrinkage.

Approach

Step 1: Understand the Basis for Night Calls

The provided assumption says:

For every 100 daytime calls (9 AM – 9 PM), 30 calls are expected during night hours (9 PM – 9 AM).

Thus, $\text{Total Night Calls} = 30\%$ of total day calls.

Total Day Call Seconds	**16463119**
Total Night Call Seconds	4938935.7

Step 2: Distribute Night Calls Across Time Buckets

Using the provided call distribution image:

Time Bucket	Count	% of Night Calls
21_22	3	10%
22_23	3	10%
23_0	2	7%
0_1	2	7%
1_2	1	3%
2_3	1	3%
3_4	1	3%
4_5	1	3%
5_6	3	10%
6_7	4	13%
7_8	4	13%
8_9	5	17%

This table was created by manually converting the given frequency distribution to percentages.

Step 3: Allocate Total Night Call Seconds to Each Time Bucket

We first calculated the **total call seconds during night** and then allocated it proportionally using the % of Night Calls.

For example:
If total night call seconds = **4,938,935.7 seconds**, then for 10% (e.g., 9–10 PM):

$$\text{Call Seconds} = 10\% \times 4{,}938{,}935.7 = 493{,}893.57$$

Apply this for each time bucket using its percentage

Step 4: Convert Call Seconds to Agent Hours

$$\text{Agent Hours} = \text{Call Seconds}/3600$$

Step 5: Account for Shrinkage

Shrinkage factor = **1.42857** (assumes 70% productivity)

$$Adjusted\ Agent\ Hours = Agent\ Hours \times 1.42857$$

Step 6: Estimate Agents Needed

Assume each agent works 1 hour per hour block (i.e., no multi-hour overlapping shifts).
Round up Adjusted Agent Hours to estimate the number of agents needed.

$$\text{Agents Needed} = \lceil \text{Adjusted Agent Hours} \rceil$$

Findings

The call seconds and agent needs for each hourly bucket were estimated as follows:

Time Bucket	Count	% of Night Calls	Call Seconds	Agent Hours	Adjusted Agent Hours	Agents Needed
21_22	3	10%	493893.57	137.1926583	195.9895119	196
22_23	3	10%	493893.57	137.1926583	195.9895119	196
23_0	2	7%	329262.38	91.46177222	130.6596746	131
0_1	2	7%	329262.38	91.46177222	130.6596746	131
1_2	1	3%	164631.19	45.73088611	65.3298373	66
2_3	1	3%	164631.19	45.73088611	65.3298373	66
3_4	1	3%	164631.19	45.73088611	65.3298373	66
4_5	1	3%	164631.19	45.73088611	65.3298373	66
5_6	3	10%	493893.57	137.1926583	195.9895119	196
6_7	4	13%	658524.76	182.9235444	261.3193492	262
7_8	4	13%	658524.76	182.9235444	261.3193492	262
8_9	5	17%	823155.95	228.6544306	326.6491865	327

Insights

- **Night shift staffing requires clear planning** even if only 30% of traffic occurs at night. Without it, SLA targets may not be met.

- Most calls between 9 PM - 11 PM and 6 AM - 9 AM; these slots need **relatively more agents**.
- Applying **shrinkage** increases the actual manpower requirement by 42.86%, highlighting the importance of factoring in non-productive time.
- The methodology allows dynamic adjustment - if actual night traffic or shrinkage varies, recalculations can be made easily.

Conclusion

The ABC Call Volume Trend Analysis project provided actionable insights into both **call durations** and **call volumes** across different hours of the day, enabling data-driven workforce planning and CX (Customer Experience) optimization.

Key takeaways include:

- **Peak call volumes** occur between **10 AM to 1 PM**, indicating the need for maximum agent availability during these hours to manage demand efficiently.
- **Longest call durations** are observed from **3 PM to 6 PM**, pointing to complex customer queries or heightened engagement levels during mid to late afternoons.
- There is a **mismatch** between peak volume and peak duration periods, emphasizing the importance of not just quantity-based but also quality-based staffing strategies.
- These findings support informed decisions for **agent scheduling**, **skill-based routing**, and **operational alignment** to improve overall service quality and reduce customer wait times.

This analytical foundation enables ABC to not only improve service operations but also better anticipate customer behavior trends during the day - fostering more personalized and effective customer interactions.

https://docs.google.com/spreadsheets/d/1-zrG4XNmDoFNr0nxWLBDBmf2NJEKgnBd/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true