

# 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

|   |            |
|---|------------|
| <b>1. Professional Background .....</b>                       | <b>1</b>   |
| <b>2. Table of Contents .....</b>                             | <b>2</b>   |
| <b>3. Real-Life Analytics - Shopping Decisions .....</b>      | <b>3</b>   |
| <b>4. Instagram User Analytics (SQL).....</b>                 | <b>6</b>   |
| <b>5. Operational Analytics &amp; Metric Spike (SQL).....</b> | <b>16</b>  |
| <b>6. Bank Loan Case Study .....</b>                          | <b>32</b>  |
| <b>7. Car Features &amp; Profitability Analysis.....</b>      | <b>50</b>  |
| <b>8. ABC Call Volume Trend Analysis.....</b>                 | <b>73</b>  |
| <b>9. IMDB Movie Analysis.....</b>                            | <b>85</b>  |
| <b>10. Hiring Process Analytics.....</b>                      | <b>105</b> |
| <b>11. Learnings &amp; Final Thoughts .....</b>               | <b>119</b> |

# 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

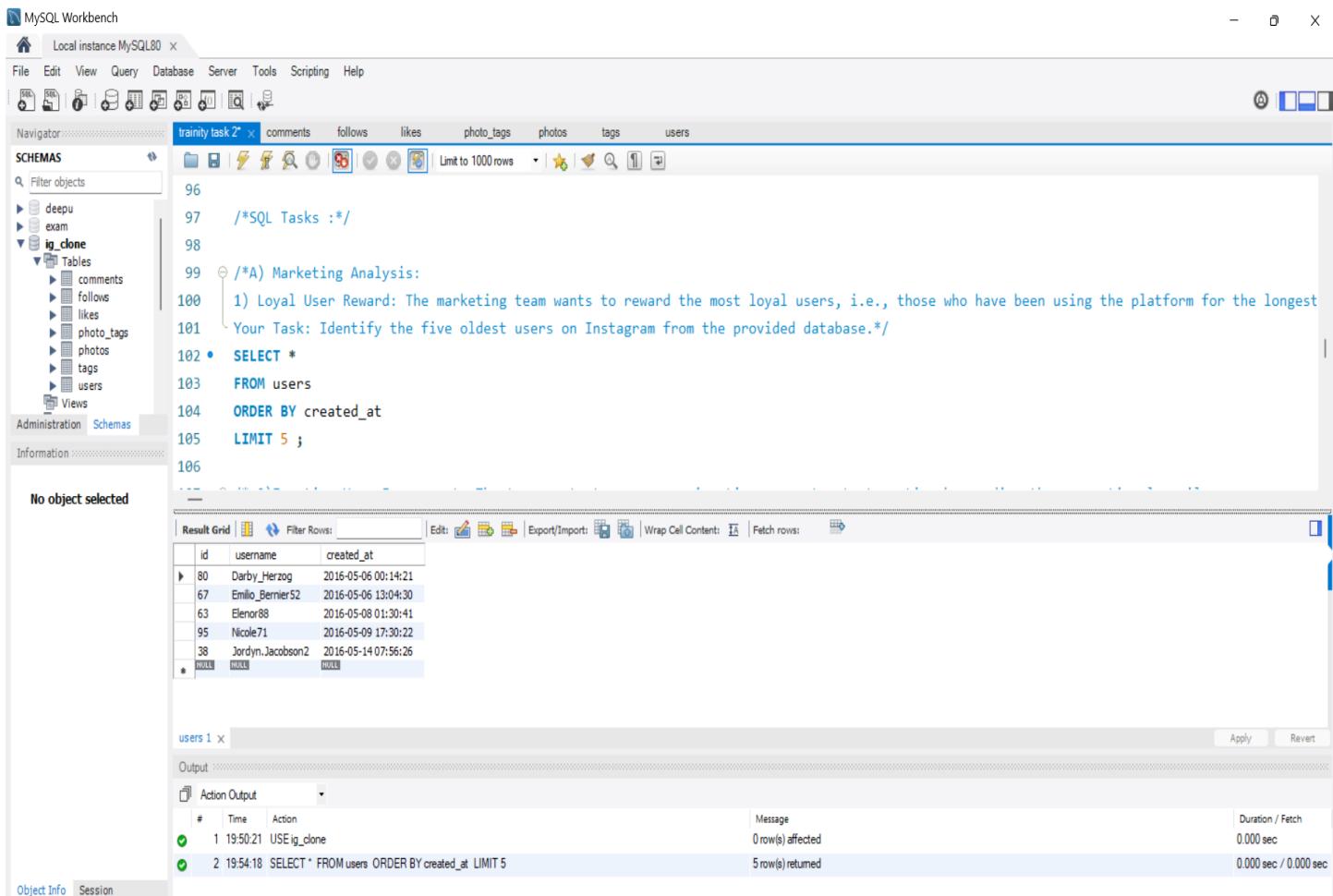
| Tool / Library  | Version | Purpose  |
|-----------------|---------|--|
| MySQL Server    | 8.x     | Hosts the relational database engine.                |
| MySQL Workbench | 8.x     | Schema design, SQL editing, output visualization.    |
| Microsoft Word  | 365     | Report drafting, embedding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



The screenshot shows the MySQL Workbench interface. The top navigation bar includes File, Edit, View, Query, Database, Server, Tools, Scripting, and Help. Below the menu is a toolbar with various icons. The main area has tabs for 'trinity task 2', 'comments', 'follows', 'likes', 'photo\_tags', 'photos', 'tags', and 'users'. On the left, the Navigator pane shows Schemas (deepu, exam, ig\_clone) and Tables (comments, follows, likes, photo\_tags, photos, tags, users). The central pane contains a SQL editor with the following code:

```

96
97     /*SQL Tasks :*/
98
99     /*A) Marketing Analysis:
100      1) Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest
101      Your Task: Identify the five oldest users on Instagram from the provided database.*/
102 • SELECT *
103   FROM users
104   ORDER BY created_at
105   LIMIT 5 ;
106

```

The Result Grid below displays the query results:

|   | id   | username         | created_at          |
|---|------|------------------|---------------------|
| ▶ | 80   | Darby_Herzog     | 2016-05-06 00:14:21 |
| ▶ | 67   | Emilio_Bernier52 | 2016-05-06 13:04:30 |
| ▶ | 63   | Elenor88         | 2016-05-08 01:30:41 |
| ▶ | 95   | Nicole71         | 2016-05-09 17:30:22 |
| ▶ | 38   | Jordyn.Jacobson2 | 2016-05-14 07:56:26 |
| * | NULL | NULL             | NULL                |

The Output pane shows the execution log:

| # | Time     | Action  | Message           | Duration / Fetch      |
|---|----------|---|-------------------|-----------------------|
| 1 | 19:50:21 | USE ig_clone                                    | 0 row(s) affected | 0.000 sec             |
| 2 | 19:54:18 | SELECT * FROM users ORDER BY created_at LIMIT 5 | 5 row(s) returned | 0.000 sec / 0.000 sec |

**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

The screenshot shows the MySQL Workbench interface with a database named 'ig\_clone'. A query titled 'trainity task 2' is run against the 'users' table. The results show 26 rows of user IDs and usernames. Below the results, the 'Output' pane shows two log entries for the executed queries.

```

105 LIMIT 5;
106
107 /* 2)Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails.
108 Your Task: Identify users who have never posted a single photo on Instagram.*/
109 • SELECT u.id, u.username
110 FROM users u
111 LEFT JOIN photos p ON u.id = p.user_id
112 WHERE p.id IS NULL;

```

| id | username           |
|----|--------------------|
| 5  | Aniya_Hackett      |
| 7  | Kassandra_Homenick |
| 14 | Jadlyn81           |
| 21 | Rocio33            |
| 24 | MaxwellHalvorson   |
| 25 | TierraTrantow      |
| 34 | Pearl7             |
| 36 | Ollie_Lederer37    |
| 41 | Mckenna17          |
| 45 | David_Osinski47    |
| 49 | MorganKassuke      |
| 53 | Linnies59          |
| 54 | Duane60            |

Result 2 | Read Only

| # | Time     | Action   | Message            | Duration / Fetch      |
|---|----------|--|--------------------|-----------------------|
| 2 | 19:54:18 | SELECT * FROM users ORDER BY created_at LIMIT 5  | 5 row(s) returned  | 0.000 sec / 0.000 sec |
| 3 | 19:55:22 | SELECT u.id, u.username FROM users u LEFT JOIN photos p ON u.id = p.user_id WHERE p.id IS NULL LIMIT ... | 26 row(s) returned | 0.000 sec / 0.000 sec |

The screenshot shows the MySQL Workbench interface with a database named 'ig\_clone'. A query titled 'trainity task 2' is run against the 'users' table. The results show a single row with the count of inactive users. Below the results, the 'Output' pane shows two log entries for the executed queries.

```

107 /* 2)Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails.
108 Your Task: Identify users who have never posted a single photo on Instagram.*/
109 • SELECT u.id, u.username
110 FROM users u
111 LEFT JOIN photos p ON u.id = p.user_id
112 WHERE p.id IS NULL;
113
114 • SELECT COUNT(*) AS inactive_user_count
115 FROM users u
116 LEFT JOIN photos p ON u.id = p.user_id
117 WHERE p.id IS NULL;
118
119 /* 3)Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins.
120 Your Task: Determine the winner of the contest and provide their details to the team.*/

```

| inactive_user_count |
|---------------------|
| 126                 |

**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

The screenshot shows the MySQL Workbench interface with a query editor and results grid.

**Query Editor:**

```

115 /* 3)Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins.
116 Your Task: Determine the winner of the contest and provide their details to the team.*/
117 • SELECT u.id AS user_id, u.username, p.id AS photo_id, COUNT(l.user_id) AS like_count
118   FROM photos p
119   JOIN likes l ON p.id = l.photo_id
120   JOIN users u ON p.user_id = u.id
121   GROUP BY p.id
122   ORDER BY like_count DESC
123   LIMIT 1;
124

```

**Result Grid:**

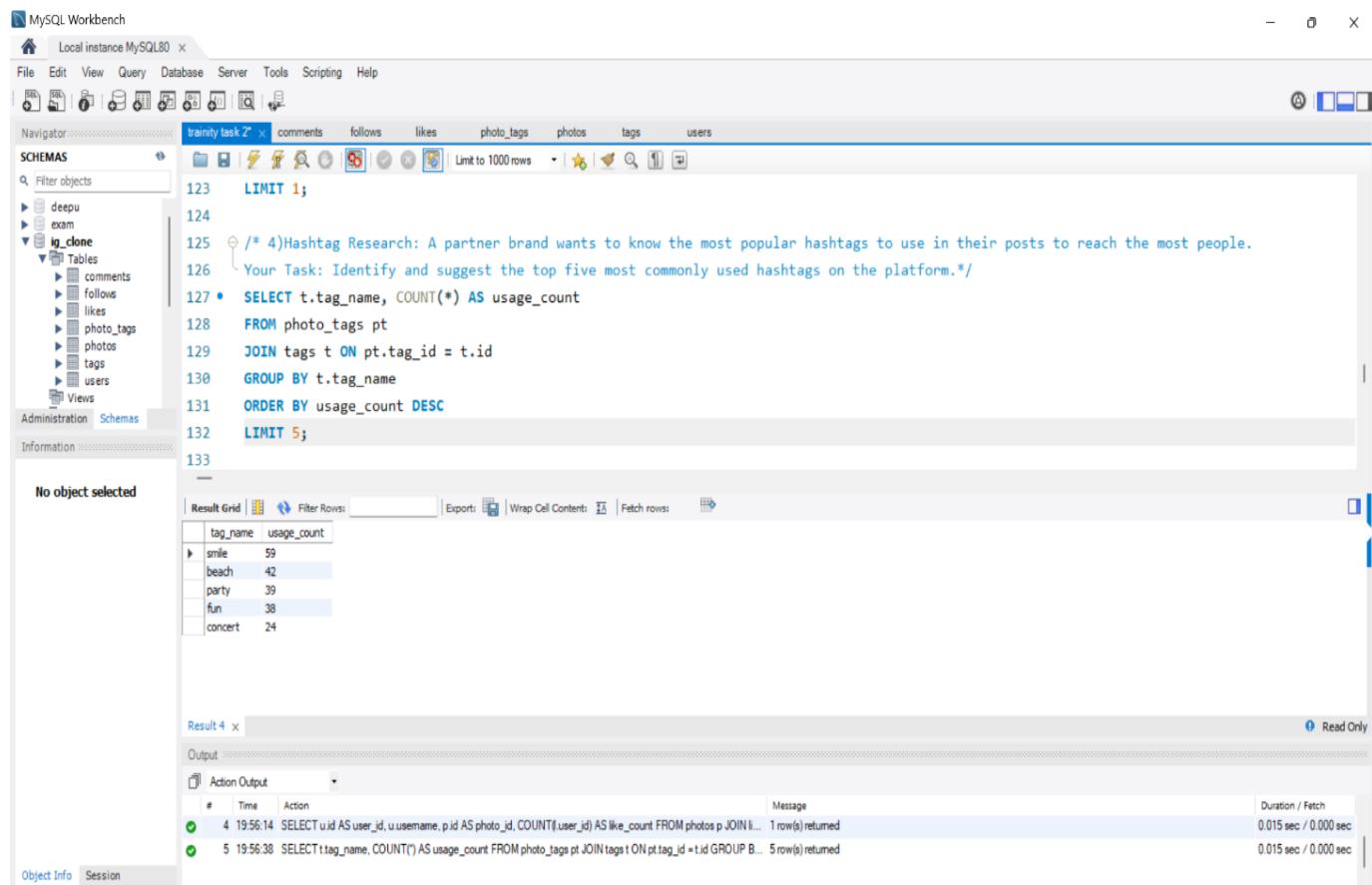
| user_id | username      | photo_id | like_count |
|---------|---------------|----------|------------|
| 52      | Zack_Kemmer93 | 145      | 48         |

**Output:**

| # | Time     | Action   | Message            | Duration / Fetch      |
|---|----------|--|--------------------|-----------------------|
| 3 | 19:55:22 | SELECT u.id, u.username FROM users u LEFT JOIN photos p ON u.id = p.user_id WHERE p.id IS NULL LIMIT ...   | 26 row(s) returned | 0.000 sec / 0.000 sec |
| 4 | 19:56:14 | SELECT u.id AS user_id, u.username, p.id AS photo_id, COUNT(l.user_id) AS like_count FROM photos p JOIN likes l ON p.id = l.photo_id GROUP BY p.id ORDER BY like_count DESC LIMIT 1; | 1 row(s) returned  | 0.015 sec / 0.000 sec |

**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



The screenshot shows the MySQL Workbench interface with a query editor window titled "trinity task 2". The code is a SQL query designed to find the top five most commonly used hashtags:

```

123 LIMIT 1;
124
125 /* 4)Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.
126 Your Task: Identify and suggest the top five most commonly used hashtags on the platform.*/
127 • SELECT t.tag_name, COUNT(*) AS usage_count
128 FROM photo_tags pt
129 JOIN tags t ON pt.tag_id = t.id
130 GROUP BY t.tag_name
131 ORDER BY usage_count DESC
132 LIMIT 5;
133

```

The result grid displays the following data:

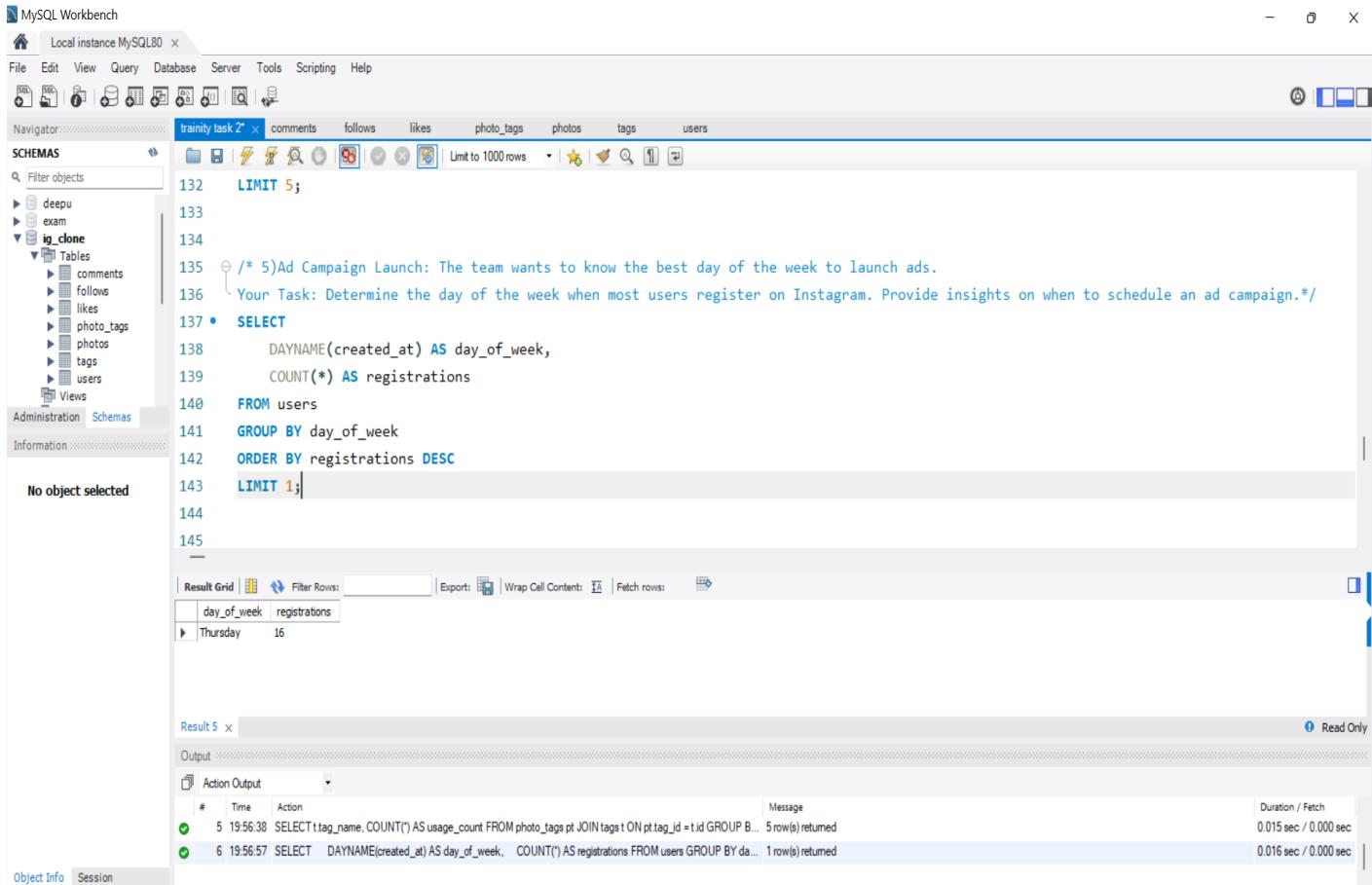
| tag_name | usage_count |
|----------|-------------|
| smile    | 59          |
| beach    | 42          |
| party    | 39          |
| fun      | 38          |
| concert  | 24          |

The output pane shows two log entries:

- # 4 19:56:14 SELECT uid AS user\_id, u.username, p.id AS photo\_id, COUNT(l.user\_id) AS like\_count FROM photos p JOIN likes l ON p.id = l.photo\_id JOIN users u ON u.id = l.user\_id WHERE p.id = 123 GROUP BY user\_id ORDER BY like\_count DESC LIMIT 1; 1 row(s) returned Duration / Fetch 0.015 sec / 0.000 sec
- # 5 19:56:38 SELECT t.tag\_name, COUNT(\*) AS usage\_count FROM photo\_tags pt JOIN tags t ON pt.tag\_id = t.id GROUP BY t.tag\_name ORDER BY usage\_count DESC LIMIT 5; 5 row(s) returned Duration / Fetch 0.015 sec / 0.000 sec

**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



The screenshot shows the MySQL Workbench interface with a query editor and results pane.

**Query Editor:**

```

132 LIMIT 5;
133
134
135 /* 5) Ad Campaign Launch: The team wants to know the best day of the week to launch ads.
136 Your Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.*/
137 • SELECT
138     DAYNAME(created_at) AS day_of_week,
139     COUNT(*) AS registrations
140     FROM users
141     GROUP BY day_of_week
142     ORDER BY registrations DESC
143     LIMIT 1;
144
145

```

**Results Grid:**

| day_of_week | registrations |
|-------------|---------------|
| Thursday    | 16            |

**Action Output:**

| # | Time     | Action   | Message           | Duration / Fetch      |
|---|----------|--|-------------------|-----------------------|
| 5 | 19:56:38 | SELECT tag_name, COUNT(*) AS usage_count FROM photo_tags pt JOIN tags t ON pt.tag_id = t.id GROUP BY tag_name ORDER BY usage_count DESC LIMIT 5; | 5 row(s) returned | 0.015 sec / 0.000 sec |
| 6 | 19:56:57 | SELECT DAYNAME(created_at) AS day_of_week, COUNT(*) AS registrations FROM users GROUP BY day_of_week ORDER BY registrations DESC LIMIT 1;        | 1 row(s) returned | 0.016 sec / 0.000 sec |

**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

The screenshot shows the MySQL Workbench interface. In the top-left corner, it says "Local instance MySQL80". The main area is a query editor titled "trinity task 2" containing the following SQL code:

```

147
148 /*1) User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.
149 Your Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by th
150 • SELECT
151     ROUND(COUNT(DISTINCT photos.id) / COUNT(DISTINCT users.id), 2) AS avg_posts_per_user
152     FROM
153         users
154     JOIN
155         photos ON users.id = photos.user_id;
156 • SELECT
157     COUNT(DISTINCT photos.id) AS total_photos,
158     COUNT(DISTINCT users.id) AS total_users,
159     ROUND(COUNT(DISTINCT photos.id) / COUNT(DISTINCT users.id), 2) AS avg_posts_per_user
160     FROM

```

Below the code is a "Result Grid" showing the output of the last query:

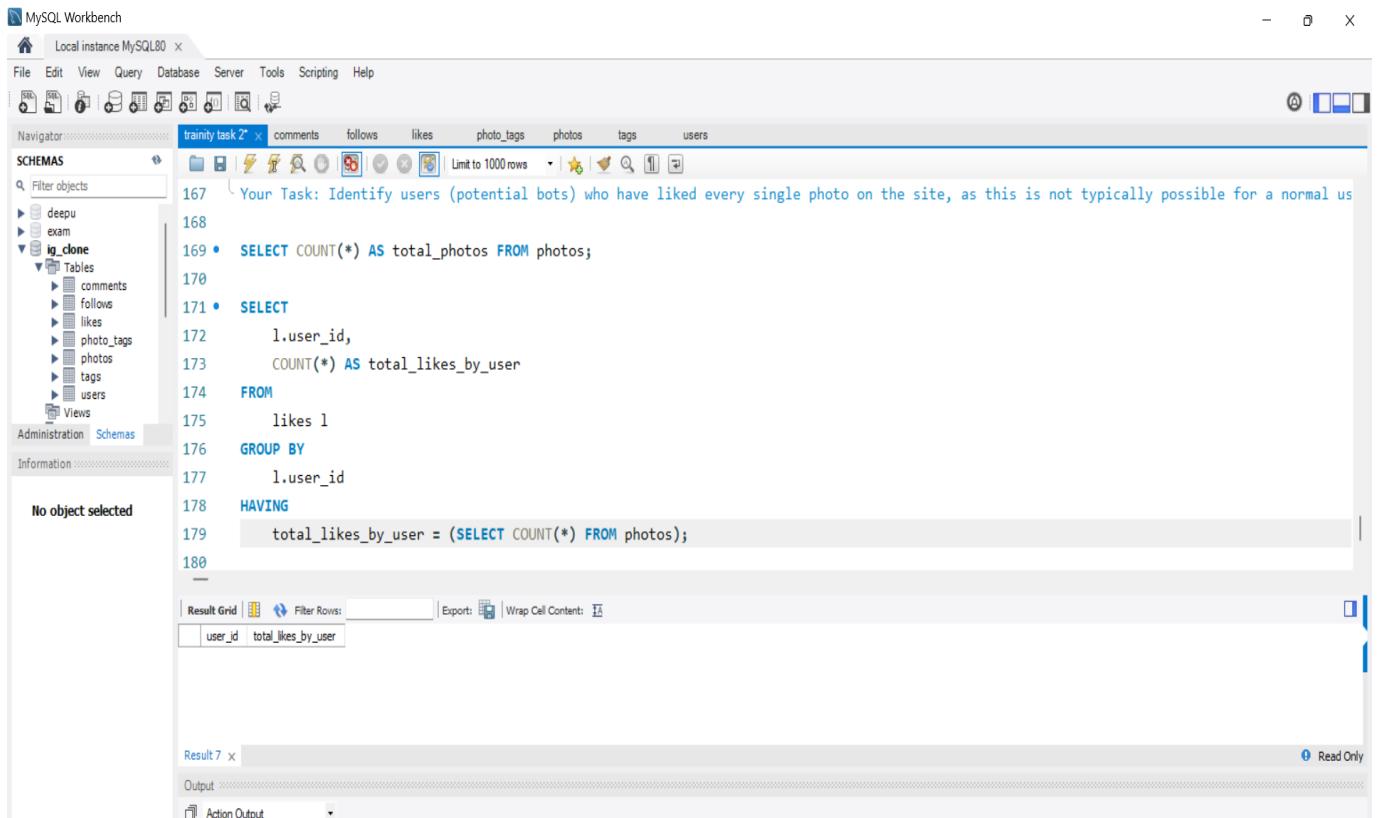
|   | avg_posts_per_user |
|---|--------------------|
| ▶ | 3.47               |

At the bottom, there is a "Result 6" tab showing the execution history:

| # | Time     | Action   | Message           | Duration / Fetch      |
|---|----------|--|-------------------|-----------------------|
| 6 | 19:56:57 | SELECT DAYNAME(created_at) AS day_of_week, COUNT(*) AS registrations FROM users GROUP BY da... | 1 row(s) returned | 0.016 sec / 0.000 sec |
| 7 | 19:57:45 | SELECT ROUND(COUNT(DISTINCT photos.id) / COUNT(DISTINCT users.id), 2) AS avg_posts_per_u...    | 1 row(s) returned | 0.000 sec / 0.000 sec |

**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



The screenshot shows the MySQL Workbench interface with a query editor window titled "trainity task 2". The code is as follows:

```

167  Your Task: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal us
168
169 •  SELECT COUNT(*) AS total_photos FROM photos;
170
171 •  SELECT
172     l.user_id,
173     COUNT(*) AS total_likes_by_user
174  FROM
175      likes l
176  GROUP BY
177      l.user_id
178  HAVING
179      total_likes_by_user = (SELECT COUNT(*) FROM photos);
180

```

The result grid shows the following data:

| user_id | total_likes_by_user |
|---------|---------------------|
|         |                     |

**Key takeaway:** While no one hit a perfect 100%, these high-percentage likers ( $\geq 90\%$ ) 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 |

|                     |       |  |
|---------------------|-------|--|
| <b>MS Word</b>      | 360   | Preparing and exporting the report               |
| <b>Google Drive</b> | 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.

The screenshot shows a SQL development environment with two panes. The top pane is titled "SQL File 10\*" and contains the following SQL code:

```

22 • ALTER TABLE job_data CHANGE ds_date ds DATE;
23
24 • SELECT * FROM job_data;
25
26 • SELECT
27     ds AS review_date,
28     COUNT(*) AS jobs_reviewed
29     FROM job_data
30     WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
31     GROUP BY ds
32     ORDER BY ds;
33
34
35

```

The bottom pane is titled "Result Grid" and displays the query results as a table:

| review_date | jobs_reviewed |
|-------------|---------------|
| 2020-11-25  | 1             |
| 2020-11-26  | 1             |
| 2020-11-27  | 1             |
| 2020-11-28  | 2             |
| 2020-11-29  | 1             |
| 2020-11-30  | 2             |

Below the Result Grid is another pane titled "Result 6" which shows the "Action Output" of the query execution:

| #  | Time     | Action  | Message           | Duration / Fetch      |
|----|----------|---|-------------------|-----------------------|
| 20 | 12:13:38 | SELECT * FROM job_data LIMIT 0, 1000  | 8 row(s) returned | 0.000 sec / 0.000 sec |
| 21 | 12:28:22 | SELECT ds AS review_date, COUNT(*) AS jobs_reviewed FROM job_data WHERE ds BETWEEN '2020-11-01' AND '2020-11-30' GROUP BY ds ORDER BY ds; | 6 row(s) returned | 0.015 sec / 0.000 sec |

**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.

SQL File 10\*

```

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

```

Output ::::::::::::

Action Output

SQL File 10\*

```

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;

```

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

| 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                     |

Result 12 \*

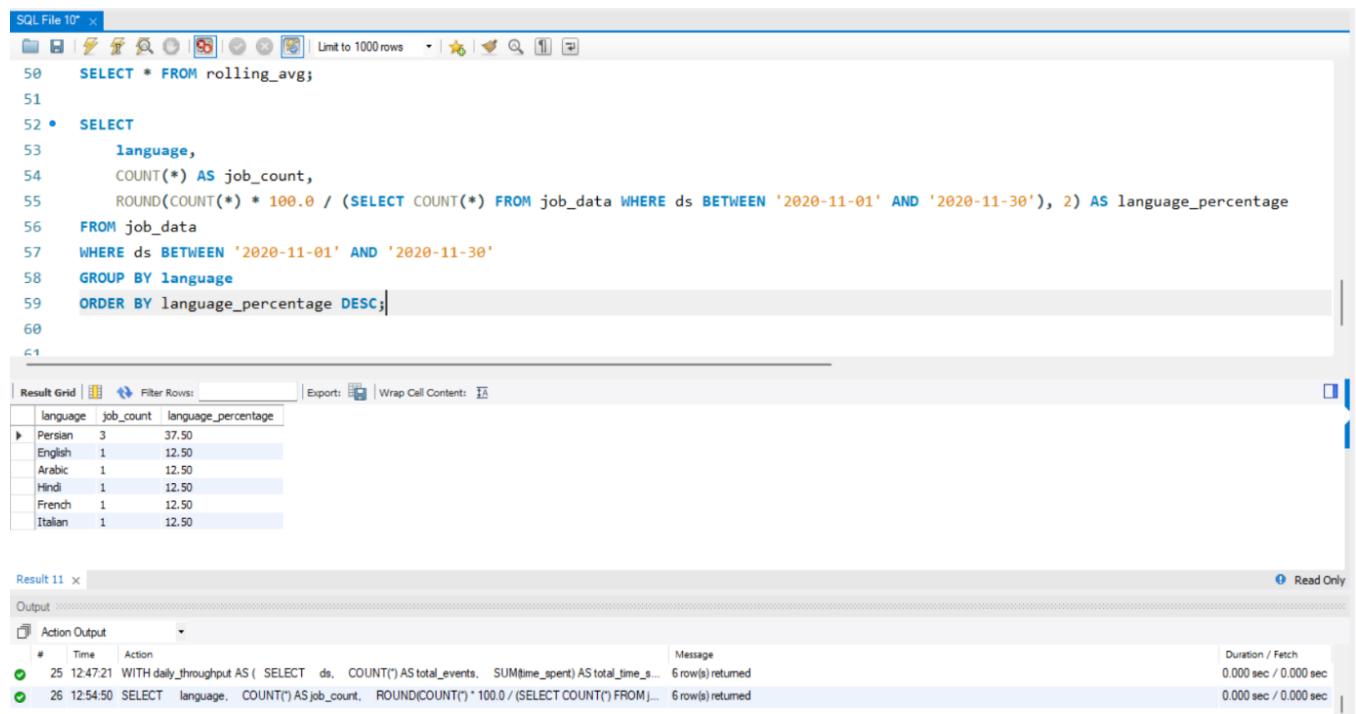
Output ::::::::::::

Read Only

**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.



The screenshot shows a SQL IDE interface with the following details:

- SQL Editor:** Contains the following SQL code:

```

50  SELECT * FROM rolling_avg;
51
52 • SELECT
53   language,
54   COUNT(*) AS job_count,
55   ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM job_data WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'), 2) AS language_percentage
56   FROM job_data
57   WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
58   GROUP BY language
59   ORDER BY language_percentage DESC;
60
61

```
- Result Grid:** Displays the output of the query as a table:

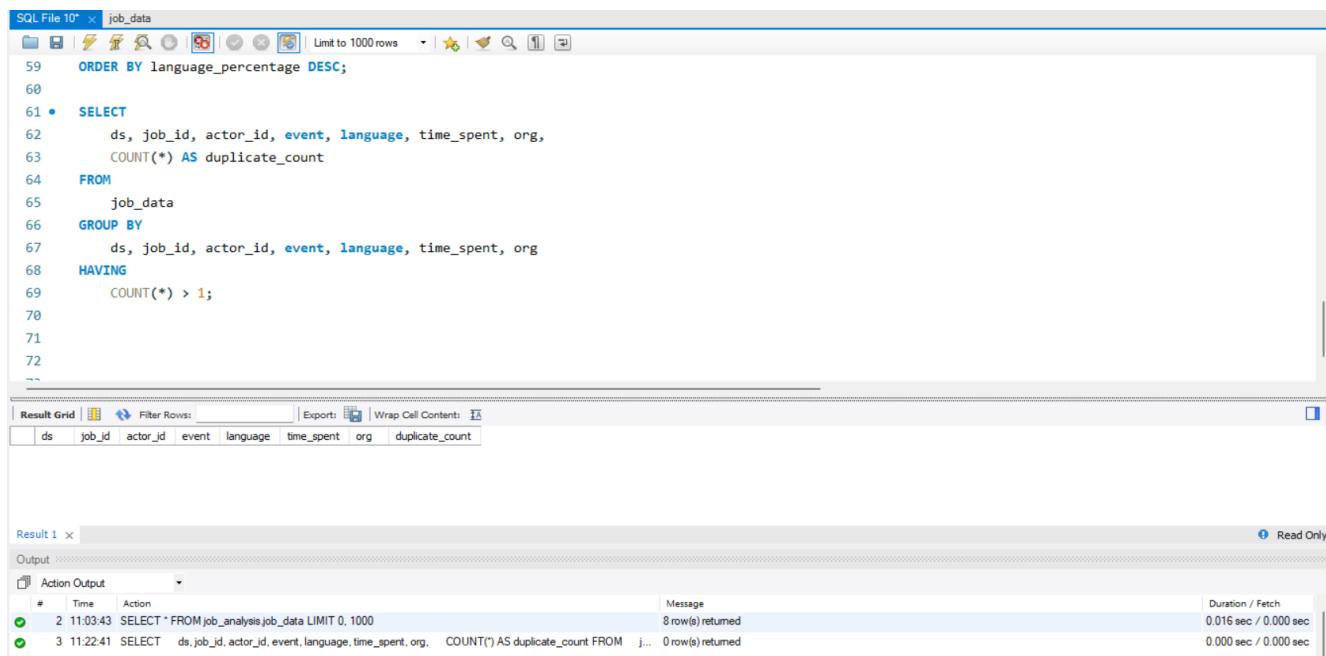
| language | job_count | language_percentage |
|----------|-----------|---------------------|
| Persian  | 3         | 37.50               |
| English  | 1         | 12.50               |
| Arabic   | 1         | 12.50               |
| Hindi    | 1         | 12.50               |
| French   | 1         | 12.50               |
| Italian  | 1         | 12.50               |
- Action Output:** Shows the history of actions taken:

| #  | Time     | Action  | Message           | Duration / Fetch      |
|----|----------|---|-------------------|-----------------------|
| 25 | 12:47:21 | WITH daily_throughput AS ( SELECT ds, COUNT() AS total_events, SUM(time_spent) AS total_time_s... | 6 row(s) returned | 0.000 sec / 0.000 sec |
| 26 | 12:54:50 | SELECT language, COUNT() AS job_count, ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM j...        | 6 row(s) returned | 0.000 sec / 0.000 sec |

**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.



The screenshot shows a SQL IDE interface with the following components:

- SQL Editor:** Displays the SQL query:
 

```

59 ORDER BY language_percentage DESC;
60
61 • SELECT
62     ds, job_id, actor_id, event, language, time_spent, org,
63     COUNT(*) AS duplicate_count
64
65     FROM
66     job_data
67     GROUP BY
68     ds, job_id, actor_id, event, language, time_spent, org
69     HAVING
70         COUNT(*) > 1;
71
72
--
```
- Result Grid:** A table with columns: ds, job\_id, actor\_id, event, language, time\_spent, org, and duplicate\_count. The grid is currently empty.
- Output:** Shows the execution log with two entries:
 

| # | Time     | Action  | Message           | Duration / Fetch      |
|---|----------|---|-------------------|-----------------------|
| 2 | 11:03:43 | SELECT * FROM job_analysis.job_data LIMIT 0, 1000   | 8 row(s) returned | 0.016 sec / 0.000 sec |
| 3 | 11:22:41 | SELECT ds, job_id, actor_id, event, language, time_spent, org, COUNT(*) AS duplicate_count FROM job_data GROUP BY ds, job_id, actor_id, event, language, time_spent, org HAVING COUNT(*) > 1; | 0 row(s) returned | 0.000 sec / 0.000 sec |

**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.

The screenshot shows a SQL editor interface with two tabs: "SQL File 10\*" and "SQL File 3\*". The "SQL File 3\*" tab is active and contains the following SQL code:

```

71  Your Task: Write an SQL query to calculate the weekly user engagement.*/
72
73 • SELECT
74     YEAR(occurred_at_dt) AS year,
75     WEEK(occurred_at_dt) AS week,
76     COUNT(DISTINCT user_id) AS active_users
77   FROM
78     events
79   GROUP BY
80     YEAR(occurred_at_dt), WEEK(occurred_at_dt)
81   ORDER BY
82     year, week;

```

Below the code is a "Result Grid" table with three columns: "year", "week", and "active\_users". The data is as follows:

| year | week | active_users |
|------|------|--------------|
| 2014 | 17   | 663          |
| 2014 | 18   | 1068         |
| 2014 | 19   | 1113         |
| 2014 | 20   | 1154         |
| 2014 | 21   | 1121         |
| 2014 | 22   | 1186         |
| 2014 | 23   | 1232         |
| 2014 | 24   | 1275         |

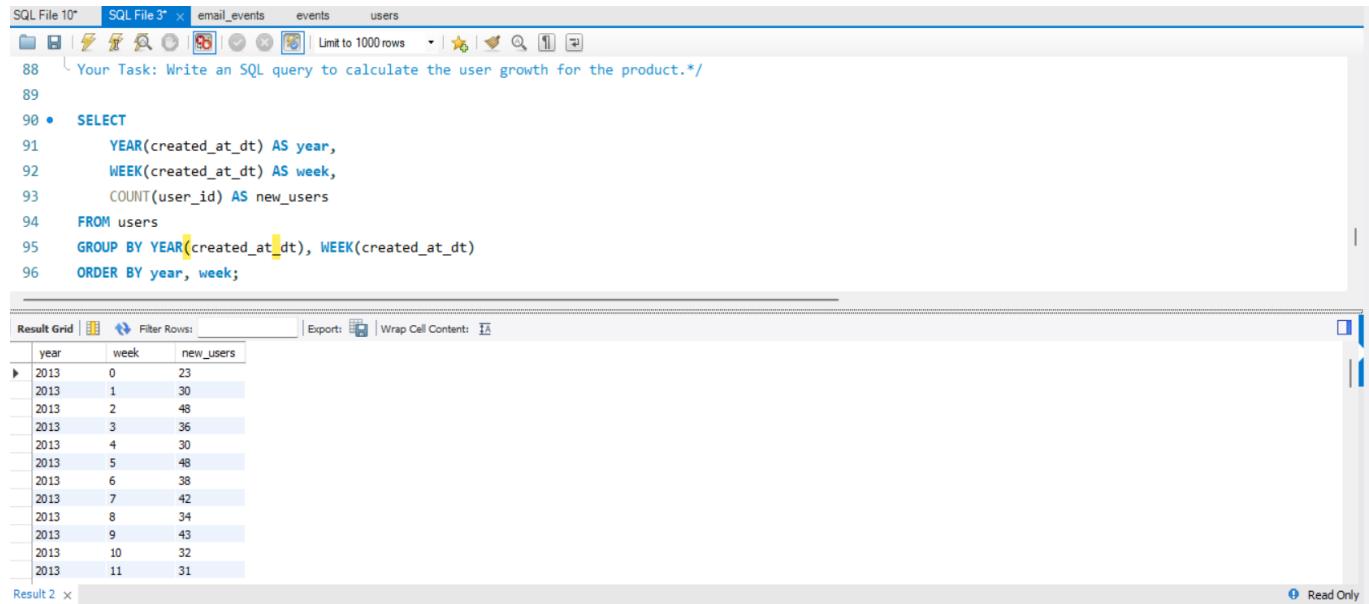
At the bottom right of the result grid, there is a "Read Only" button.

### Insights:

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



The screenshot shows a SQL editor interface with two tabs: "SQL File 10\*" and "SQL File 3\*". The "SQL File 3\*" tab is active and contains the following SQL code:

```

88  Your Task: Write an SQL query to calculate the user growth for the product.*/
89
90 • SELECT
91     YEAR(created_at_dt) AS year,
92     WEEK(created_at_dt) AS week,
93     COUNT(user_id) AS new_users
94 FROM users
95 GROUP BY YEAR(created_at_dt), WEEK(created_at_dt)
96 ORDER BY year, week;

```

Below the code is a "Result Grid" table with three columns: "year", "week", and "new\_users". The data is as follows:

| year | week | new_users |
|------|------|-----------|
| 2013 | 0    | 23        |
| 2013 | 1    | 30        |
| 2013 | 2    | 48        |
| 2013 | 3    | 36        |
| 2013 | 4    | 30        |
| 2013 | 5    | 48        |
| 2013 | 6    | 38        |
| 2013 | 7    | 42        |
| 2013 | 8    | 34        |
| 2013 | 9    | 43        |
| 2013 | 10   | 32        |
| 2013 | 11   | 31        |

### 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.

```

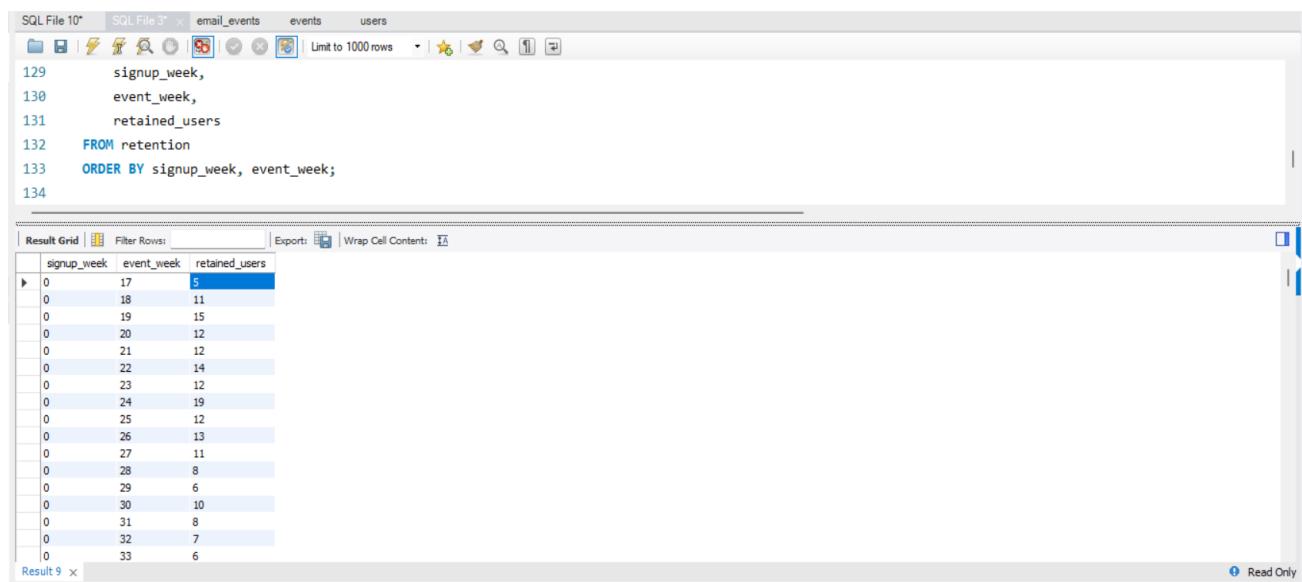
• WITH user_signup_week AS (
    SELECT
        user_id,
        MIN(DATE(activated_at_dt)) AS signup_date,
        WEEK(MIN(DATE(activated_at_dt))) AS signup_week
    FROM users
    GROUP BY user_id
),
user_events AS (
    SELECT
        user_id,
        WEEK(DATE(occurred_at_dt)) AS event_week,
        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
)

```

```

117     ),
118     retention AS (
119         SELECT
120             s.signup_week,
121             e.event_week,
122             COUNT(DISTINCT e.user_id) AS retained_users
123         FROM user_signup_week s
124         JOIN user_events e ON s.user_id = e.user_id
125         WHERE e.event_week >= s.signup_week
126         GROUP BY s.signup_week, e.event_week
127     )
128     SELECT
129         signup_week,
130         event_week,
131         retained_users
132     FROM retention
133     ORDER BY signup_week, event_week;
134
135 /*Weekly Engagement Per Device:
136 Objective: Measure the activeness of users on a weekly basis per device.
137 Your Task: Write an SQL query to calculate the weekly engagement per device.*/

```



The screenshot shows a SQL query results window. At the top, there are tabs for "SQL File 10\*", "SQL File 3\*", "email\_events", "events", and "users". Below the tabs is a toolbar with icons for file operations, search, and export. A dropdown menu says "Limit to 1000 rows". The main area contains the following SQL code:

```
129     signup_week,
130     event_week,
131     retained_users
132 FROM retention
133 ORDER BY signup_week, event_week;
134
```

Below the code is a "Result Grid" table with three columns: "signup\_week", "event\_week", and "retained\_users". The data shows a decline in retained users over time, starting at week 17 with 5 users and ending at week 33 with 6 users. The table has 15 rows.

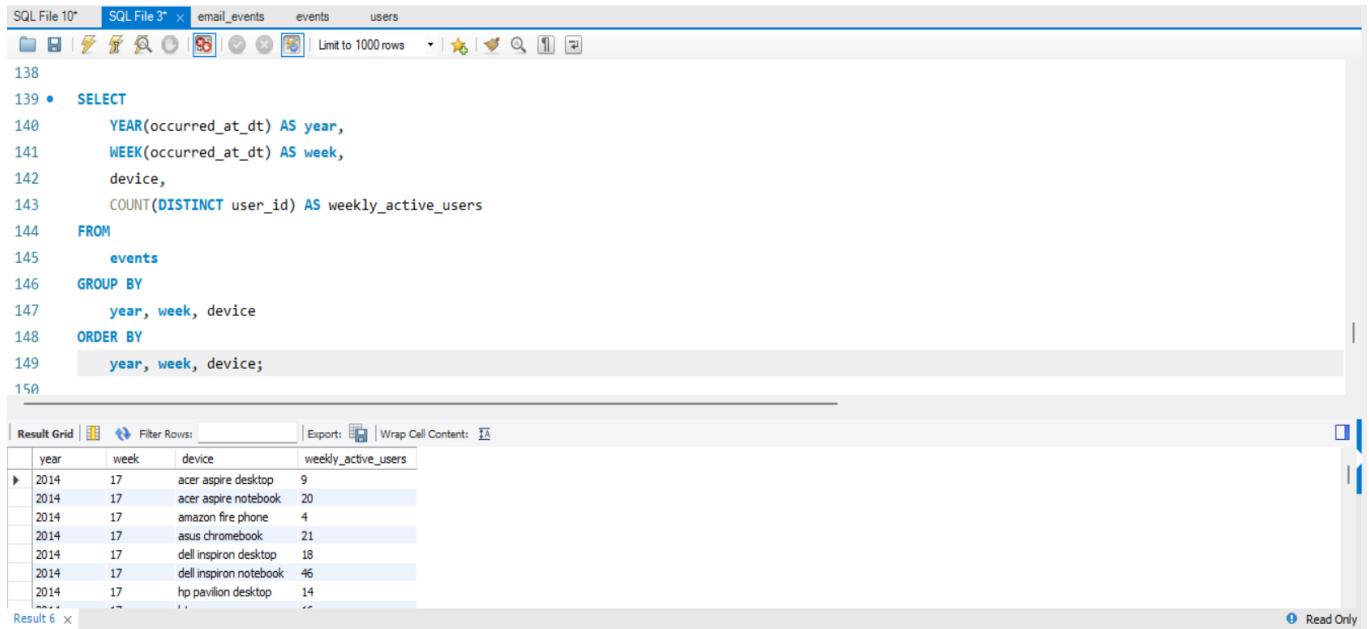
| signup_week | event_week | retained_users |
|-------------|------------|----------------|
| 0           | 17         | 5              |
| 0           | 18         | 11             |
| 0           | 19         | 15             |
| 0           | 20         | 12             |
| 0           | 21         | 12             |
| 0           | 22         | 14             |
| 0           | 23         | 12             |
| 0           | 24         | 19             |
| 0           | 25         | 12             |
| 0           | 26         | 13             |
| 0           | 27         | 11             |
| 0           | 28         | 8              |
| 0           | 29         | 6              |
| 0           | 30         | 10             |
| 0           | 31         | 8              |
| 0           | 32         | 7              |
| 0           | 33         | 6              |

At the bottom left, it says "Result 9" and "Read Only".

**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.



The screenshot shows a SQL IDE interface with two tabs at the top: "SQL File 10\*" and "SQL File 3\*". The "SQL File 10\*" tab contains the following SQL code:

```

138
139 •  SELECT
140     YEAR(occurred_at_dt) AS year,
141     WEEK(occurred_at_dt) AS week,
142     device,
143     COUNT(DISTINCT user_id) AS weekly_active_users
144   FROM
145     events
146   GROUP BY
147     year, week, device
148   ORDER BY
149     year, week, device;
150

```

The "Result Grid" tab displays the query results in a table:

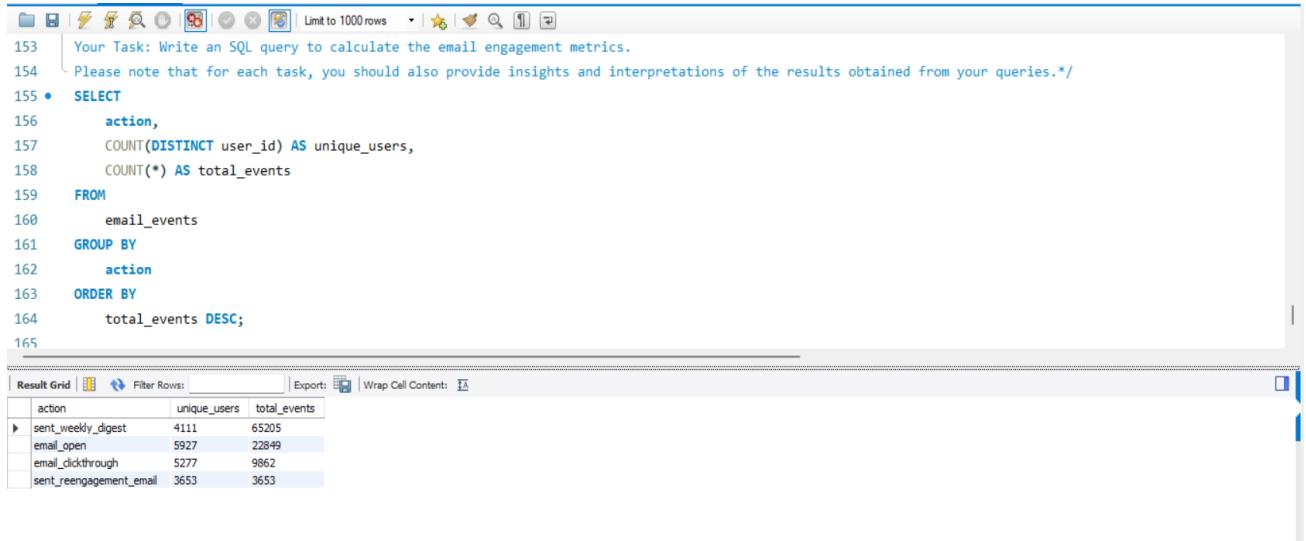
| year | week | device                 | weekly_active_users |
|------|------|------------------------|---------------------|
| 2014 | 17   | acer aspire desktop    | 9                   |
| 2014 | 17   | acer aspire notebook   | 20                  |
| 2014 | 17   | amazon fire phone      | 4                   |
| 2014 | 17   | asus chromebook        | 21                  |
| 2014 | 17   | dell inspiron desktop  | 18                  |
| 2014 | 17   | dell inspiron notebook | 46                  |
| 2014 | 17   | hp pavilion desktop    | 14                  |
| ...  | ...  | ...                    | ...                 |

Below the table, it says "Result 6" and "Read Only".

**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.



The screenshot shows a SQL editor interface with the following content:

```

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:**

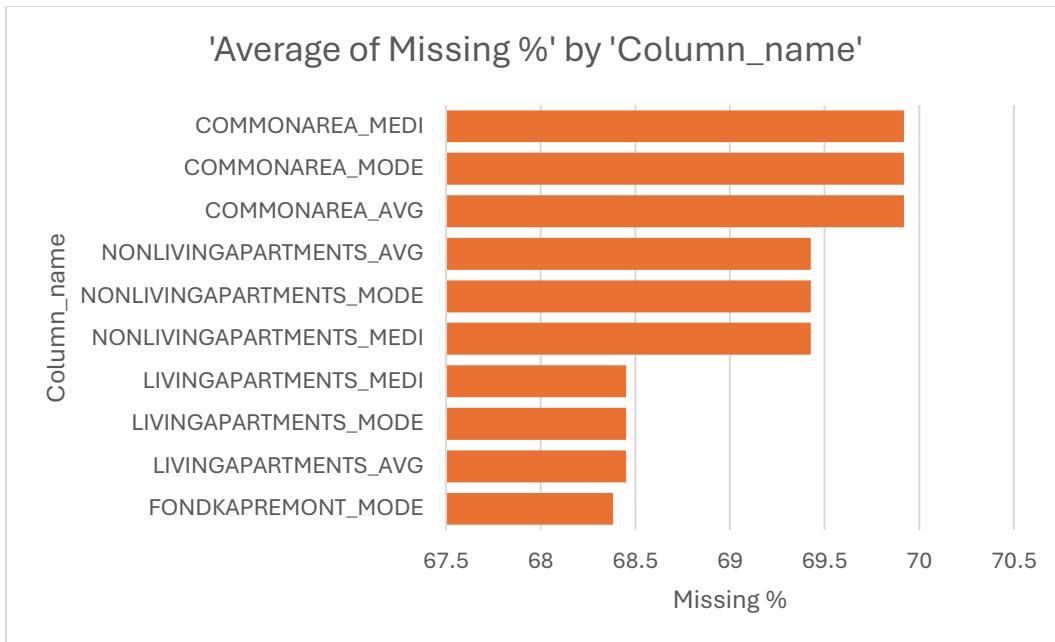
The analysis of missing data was conducted using built-in Excel functions such as:

- 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   |
|-------------------------|---------------------|--|
| <b>AMT_INCOME_TOTAL</b> | AMT_INCOME_TOTAL_CL | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |
| <b>AMT_ANNUITY</b>      | AMT_ANNUITY_CL      | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |
| <b>AMT_GOODS_PRICE</b>  | AMT_GOODS_PRICE_CL  | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |

|                                   |                               |  |
|-----------------------------------|-------------------------------|--|
| <b>AMT_CREDIT</b>                 | AMT_CREDIT_CL                 | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |
| <b>REGION_POPULATION_RELATIVE</b> | REGION_POPULATION_RELATIVE_CL | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |
| <b>EXT_SOURCE_3</b>               | EXT_SOURCE_3_CL               | Capped between 5 <sup>th</sup> and 95 <sup>th</sup> percentile |
| <b>EXT_SOURCE_2</b>               | <i>(No outliers found)</i>    | Kept as-is   |

### Treatment Strategy:

- Values below the 5th percentile were **clipped** to the 5<sup>th</sup> and 95<sup>th</sup> 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 **5<sup>th</sup> and 95<sup>th</sup> 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:

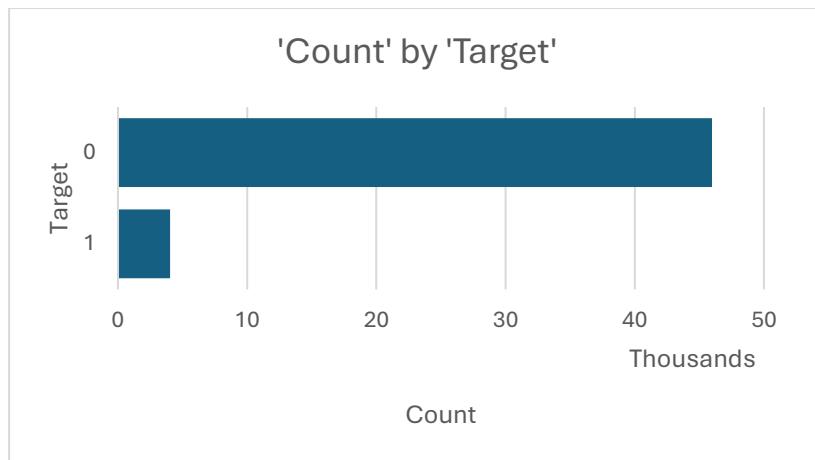
$$\text{Percent} = (\text{Total Count}/\text{Class Count}) \times 100$$

### Results:

| <b>Target Value</b> | <b>Count</b>  | <b>Percentage (%)</b> |
|---------------------|---------------|-----------------------|
| 0 (No Default)      | 45,973        | 91.95%                |
| 1 (Default)         | 4,026         | 8.05%                 |
| <b>Total</b>        | <b>49,999</b> | <b>100%</b>           |

### 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.

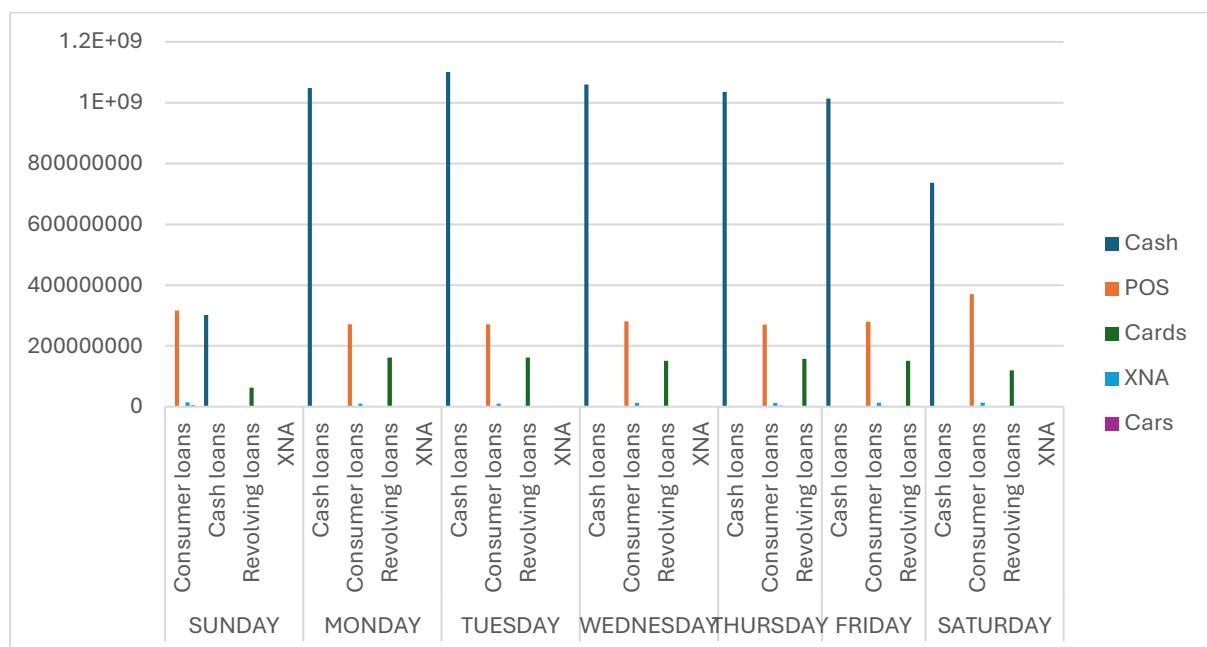
### 1. Breakdown by Weekday of Application & Loan Portfolio Type

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

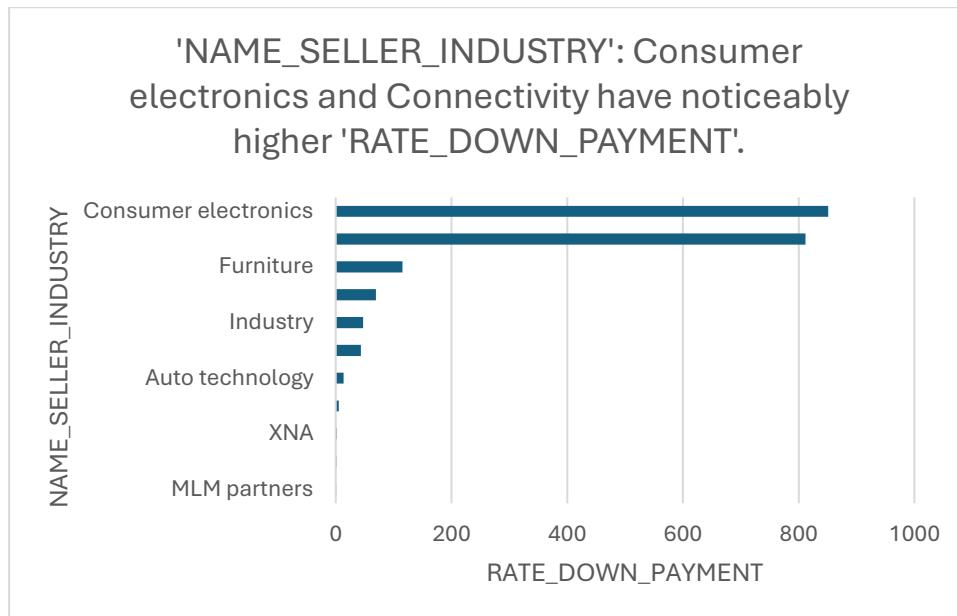
|                       |                 |                         |                         |                            |                         |                       |                        |
|-----------------------|-----------------|-------------------------|-------------------------|----------------------------|-------------------------|--------------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

## 2. Seller Industry vs. Rate of Down Payment

| Seller Industry             |  | Rate Down Payment |
|-----------------------------|--|-------------------|
| <b>Consumer electronics</b> |  | <b>851.08</b>     |
| <b>Connectivity</b>         |  | <b>811.65</b>     |
| <b>Furniture</b>            |  | 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.



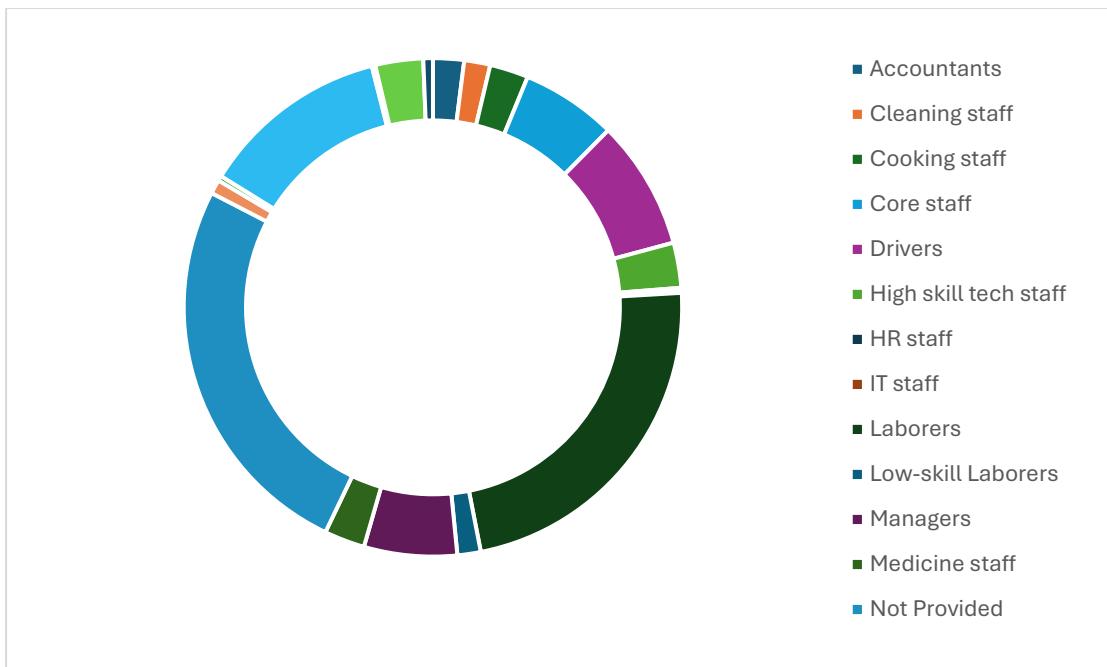
### 3. Occupation Type vs. Default Rate

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 |
|-----------------|---------------|-------|
| <i>Laborers</i> | 920           | 8952  |
| <i>Managers</i> | 243           | 3489  |
| <i>IT staff</i> | 4             | 80    |

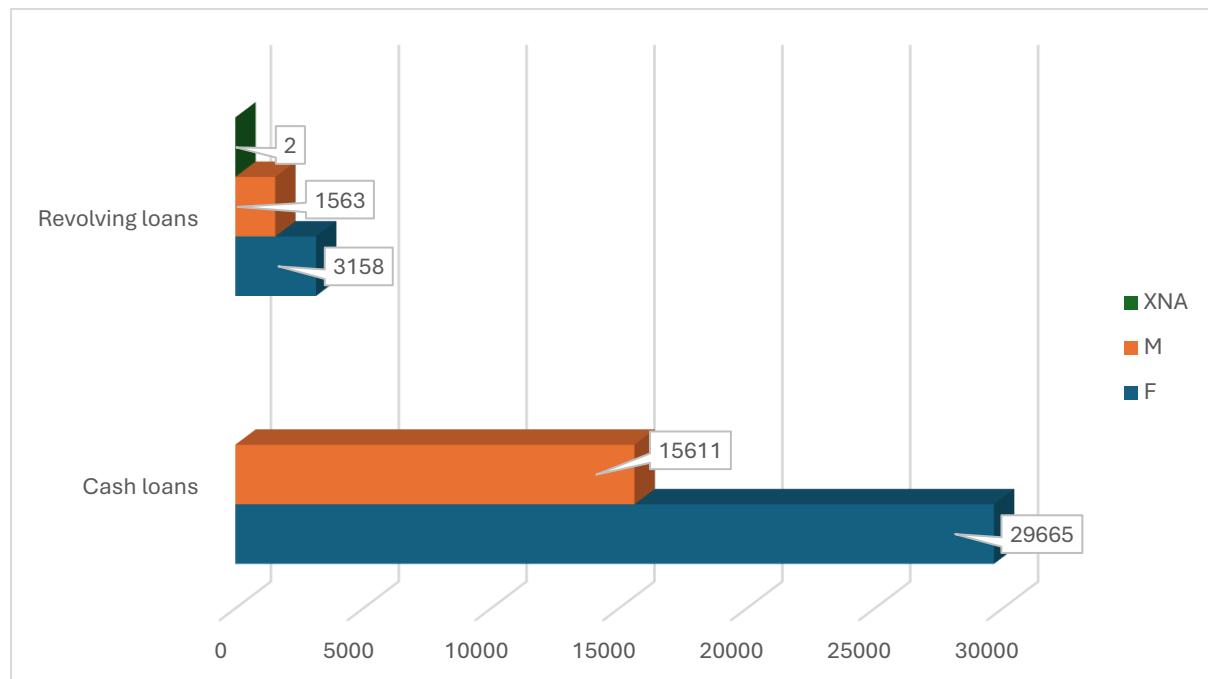
**Implication:** Occupation is a strong proxy for repayment behavior. It should be considered in credit risk models.



#### 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.

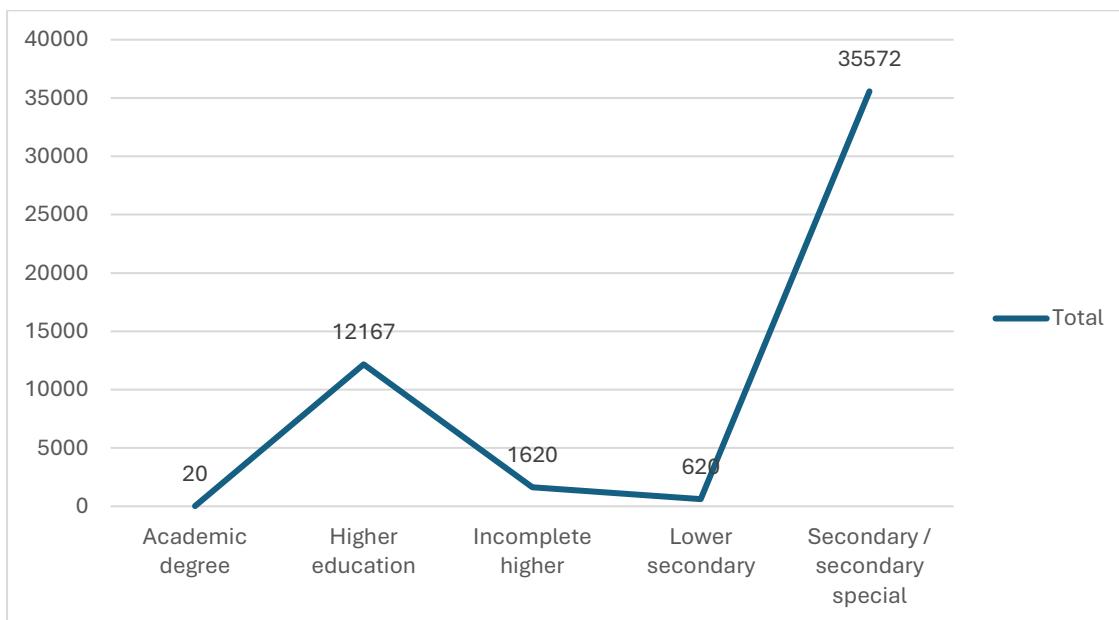
**Implication:** Loan marketing and communication can be gender-personalized, particularly for financial literacy products.



## 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.



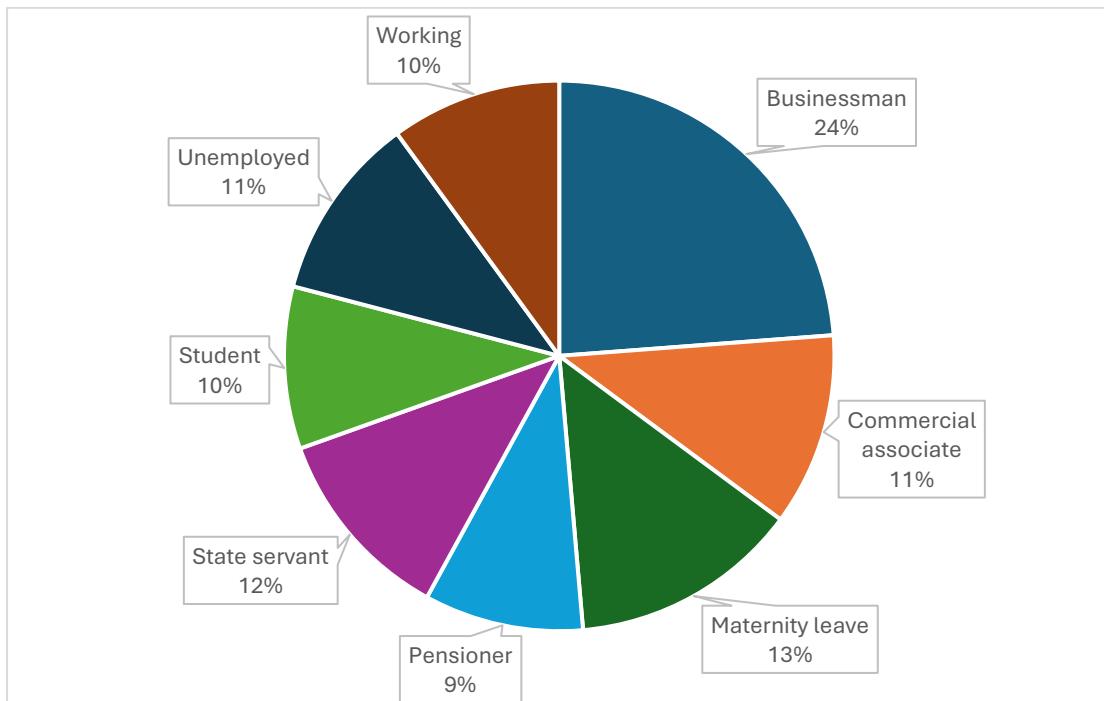
## 6. Employment Type and Loan Size

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 ( $\approx 585,408$ ).

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

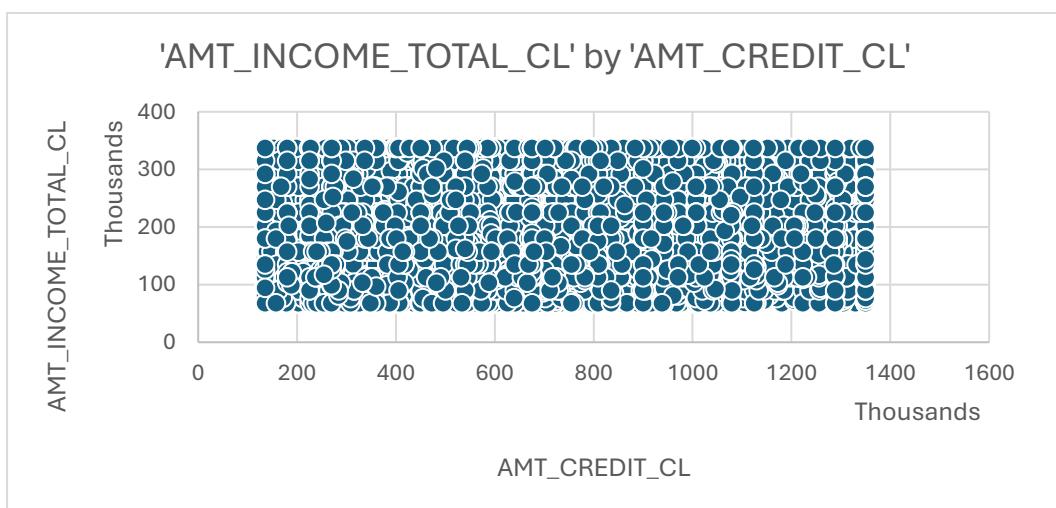


## 7. Credit Correlation Plots

Two scatterplots were examined:

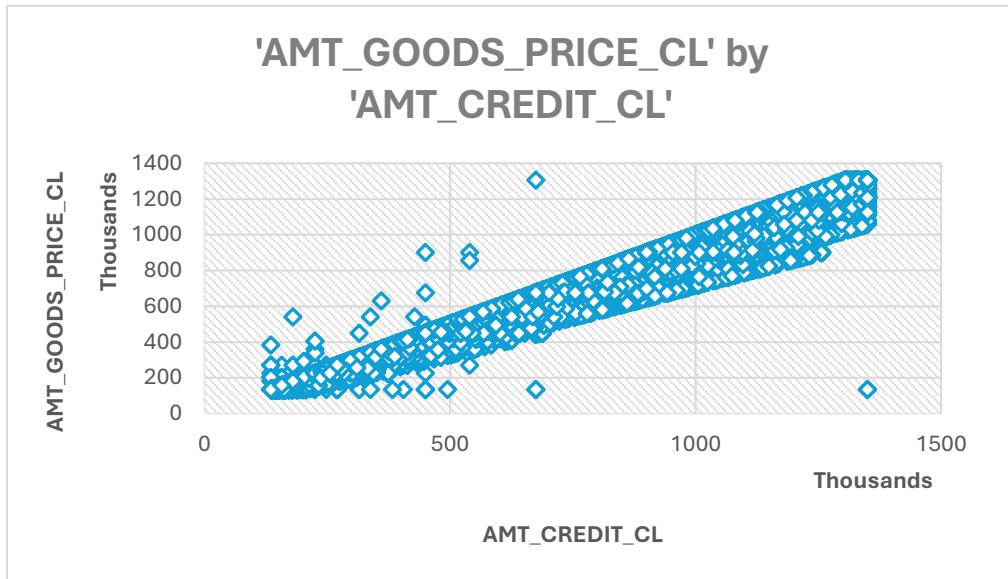
### 1. AMT\_INCOME\_TOTAL\_CL vs AMT\_CREDIT\_CL

- 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.

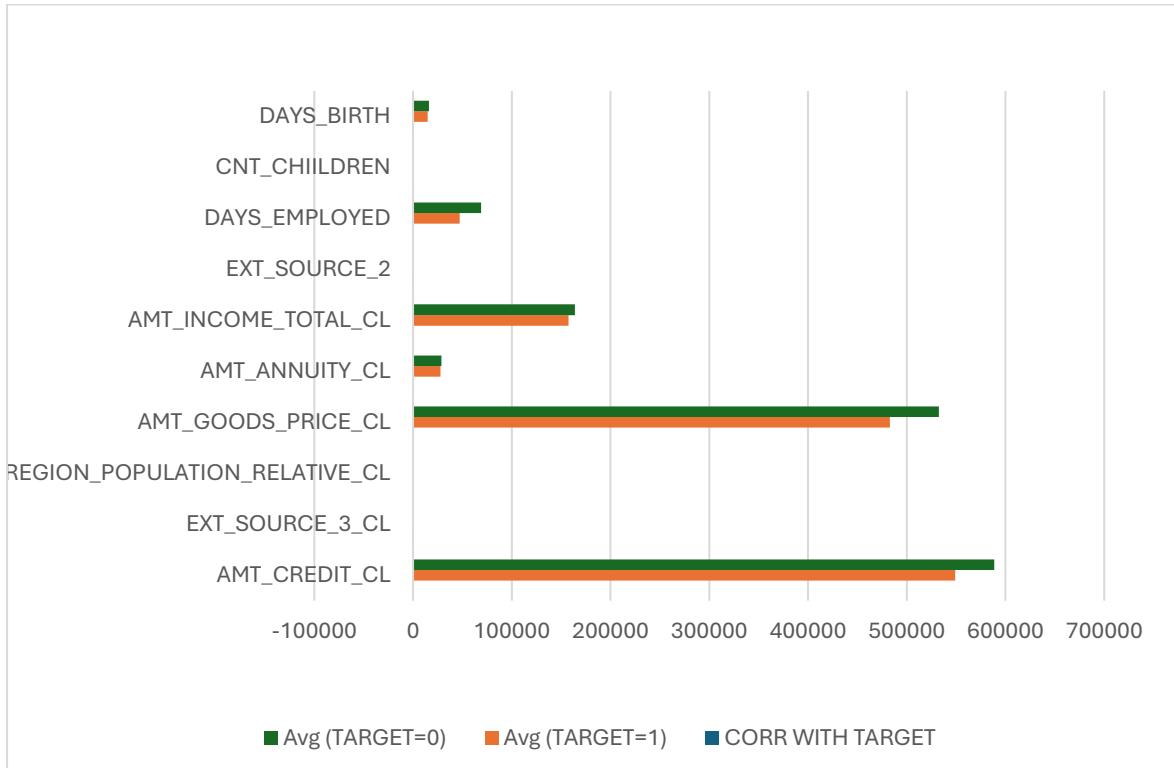
### Results Summary:

| Feature                              | Correlation with Target | Avg (Defaulters) | Avg (Non-Defaulters) | Difference |
|--------------------------------------|-------------------------|------------------|----------------------|------------|
| <b>EXT_SOURCE_2</b>                  | -0.1584                 | 0.4115           | 0.5228               | 0.1113     |
| <b>EXT_SOURCE_3_CL</b>               | -0.1543                 | 0.4321           | 0.5255               | 0.0934     |
| <b>DAYS_BIRTH</b>                    | -0.0768                 | 14,890           | 16,121               | 1,231      |
| <b>AMT_GOODS_PRICE_CL</b>            | -0.0407                 | 483,019          | 532,551              | 49,532     |
| <b>DAYS_EMPLOYED</b>                 | -0.0425                 | 47,217           | 68,907               | 21,690     |
| <b>REGION_POPULATION_RELATIVE_CL</b> | -0.0385                 | 0.0187           | 0.0203               | 0.0016     |
| <b>AMT_CREDIT_CL</b>                 | -0.0303                 | 548,940          | 588,602              | 39,662     |
| <b>AMT_INCOME_TOTAL_CL</b>           | -0.0245                 | 157,283          | 163,844              | 6,561      |
| <b>AMT_ANNUITY_CL</b>                | -0.0161                 | 27,837           | 28,608               | 771        |
| <b>CNT_CHILDREN</b>                  | <b>+0.0264</b>          | 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   |
|--|---|
| <b>Microsoft Excel (Office 365)</b>          | Primary tool for data cleaning, transformation, pivoting, regression, visualization, and dashboarding |
| <b>Excel Add-ins (Data Analysis ToolPak)</b> | Used for regression modeling and correlation analysis   |
| <b>Formulas Used</b>                         | IF(), SUMIF(), AVERAGEIF(), COUNTIF(), CORREL(), TEXT(), and others                                   |
| <b>Interactive Tools</b>                     | 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):

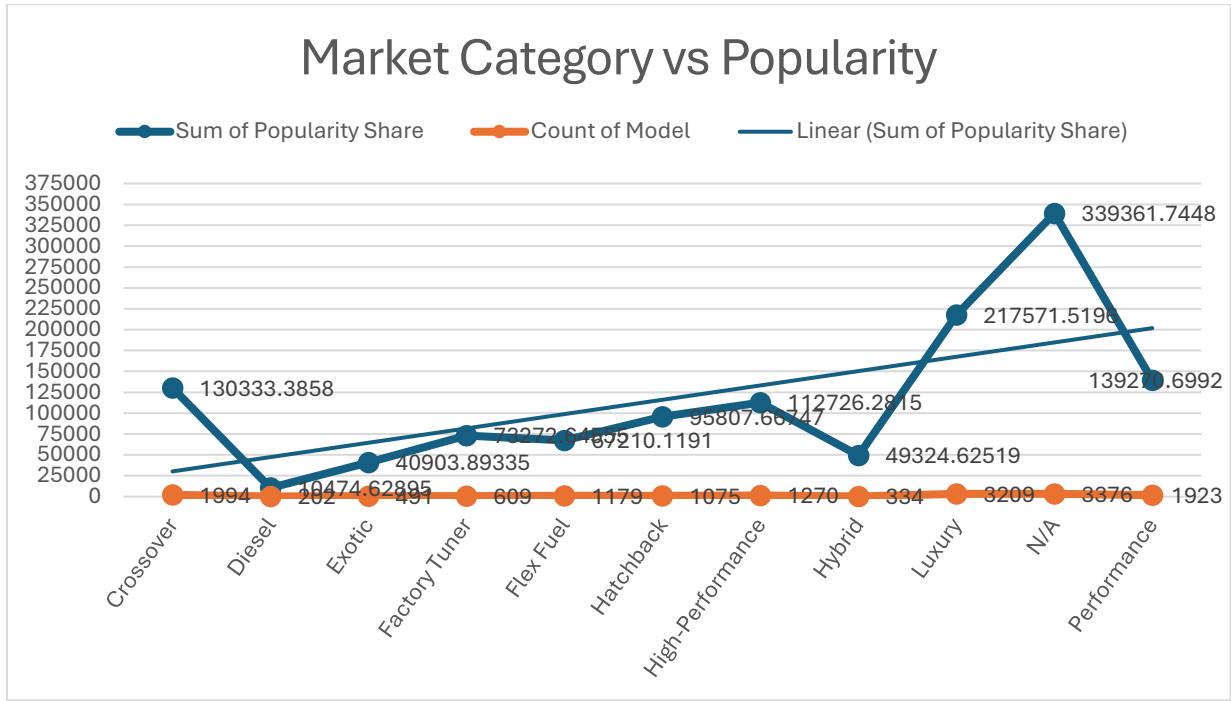
| Row Labels         | Sum of Popularity Share | Count of Model |
|--------------------|-------------------------|----------------|
| Crossover          | 130333.3858             | 1994           |
| Diesel             | 10474.62895             | 202            |
| Exotic             | 40903.89335             | 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           |
| <b>Grand Total</b> | <b>1276257.21</b>       | <b>15662</b>   |

**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:

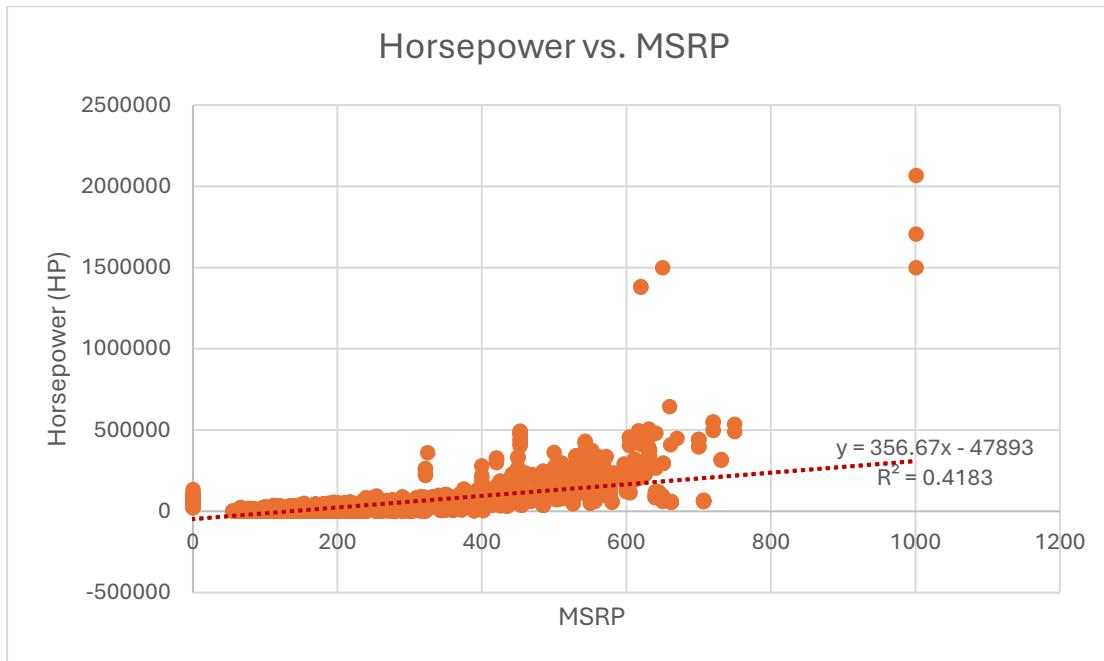
$$y = 356.67x - 47893$$

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

- The **R<sup>2</sup> 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  $\leq 0$ ).
- 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.

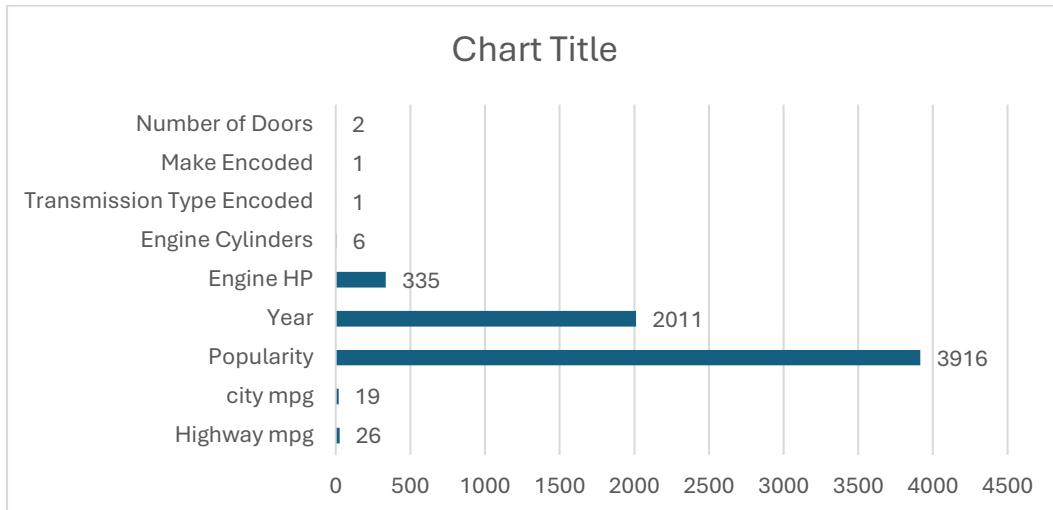
### Regression Output

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

| Feature                            | Coefficient |
|------------------------------------|-------------|
| <b>Popularity</b>                  | 3916        |
| <b>Year</b>                        | 2011        |
| <b>Engine Horsepower (HP)</b>      | 335         |
| <b>Highway MPG</b>                 | 26          |
| <b>City MPG</b>                    | 19          |
| <b>Engine Cylinders</b>            | 6           |
| <b>Transmission Type (Encoded)</b> | 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

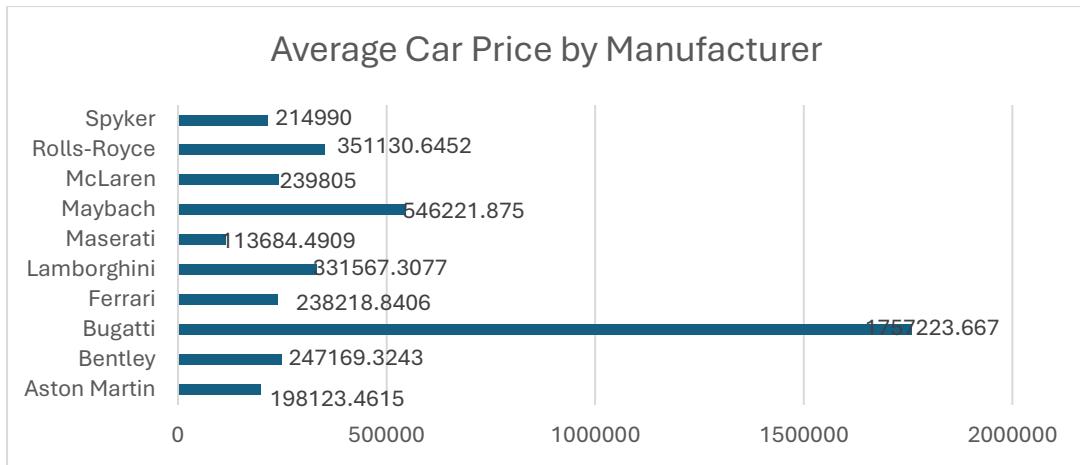
A pivot table was created to calculate the **average MSRP (Manufacturer's Suggested Retail Price)** for each car brand. The key findings are:

| 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

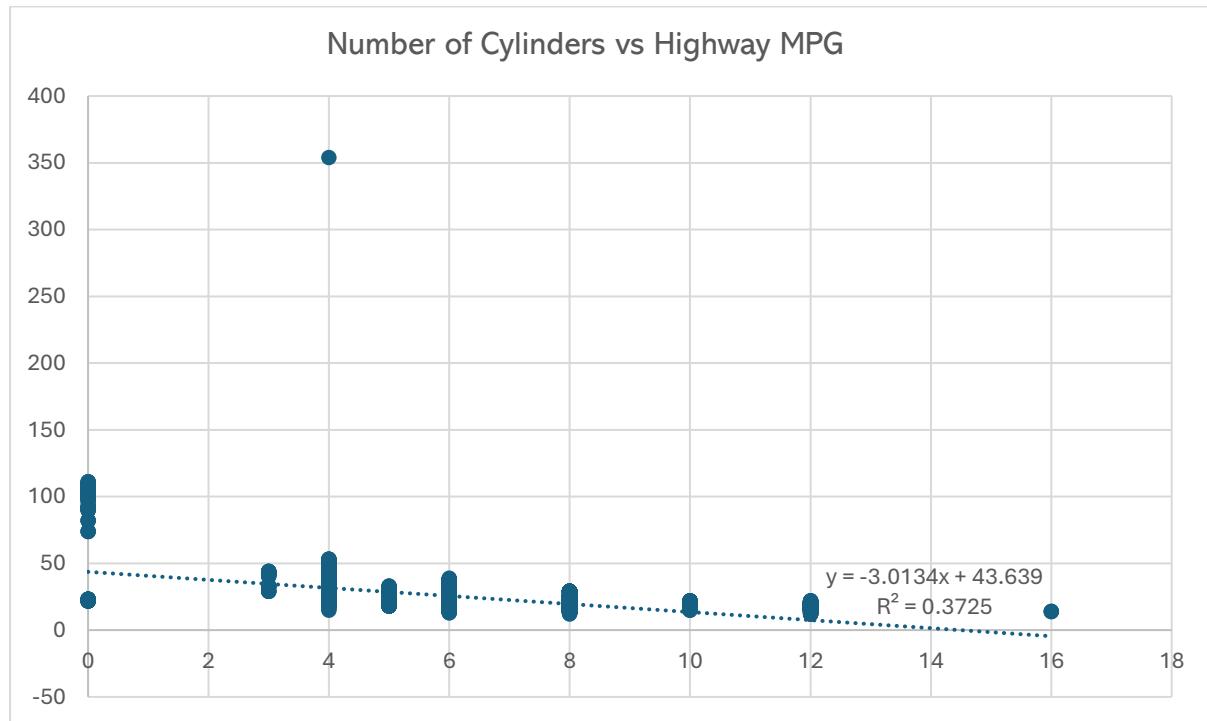
A scatter plot was created with:

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

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

$$y = -3.0134x + 43.639$$

and  $R^2 = 0.3725$



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**.

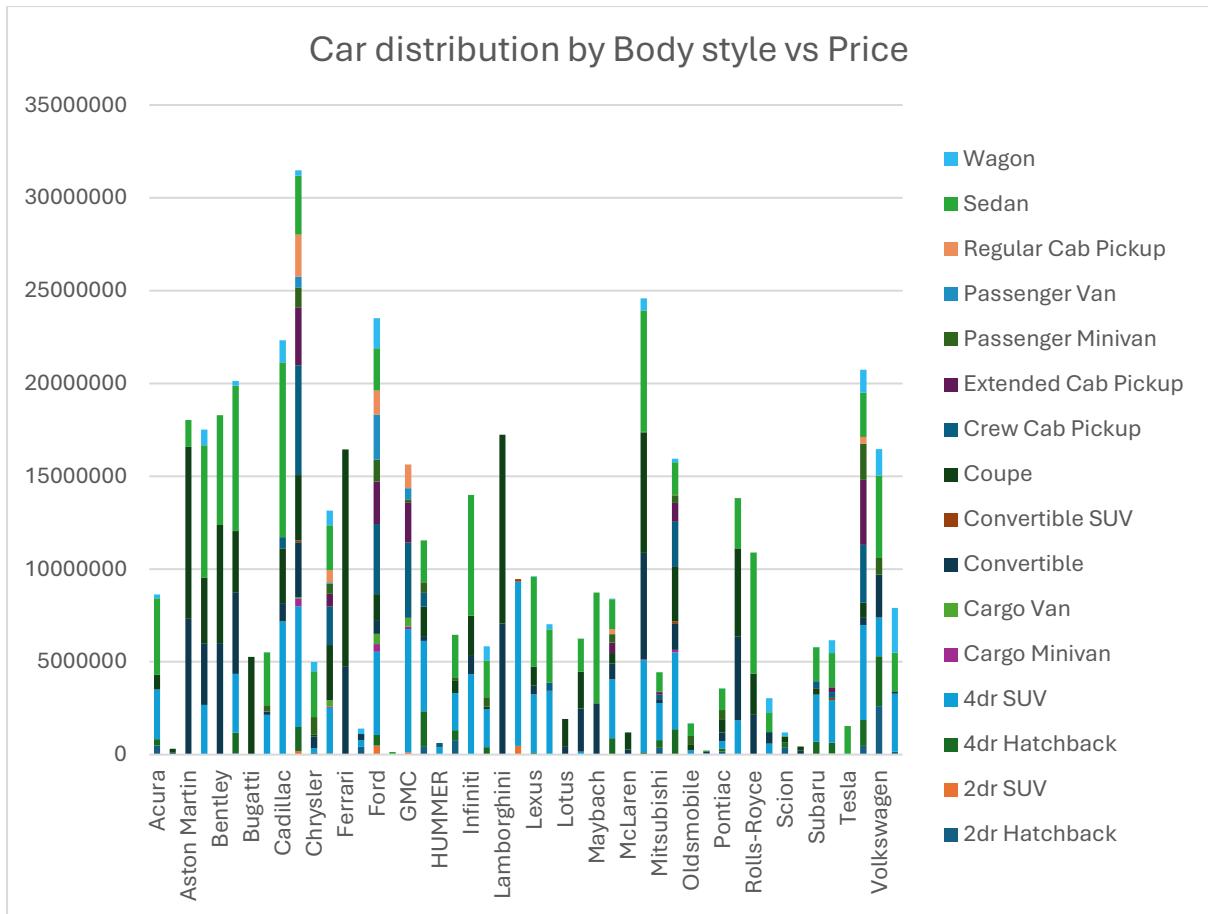
#### Key Observations:

- **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.

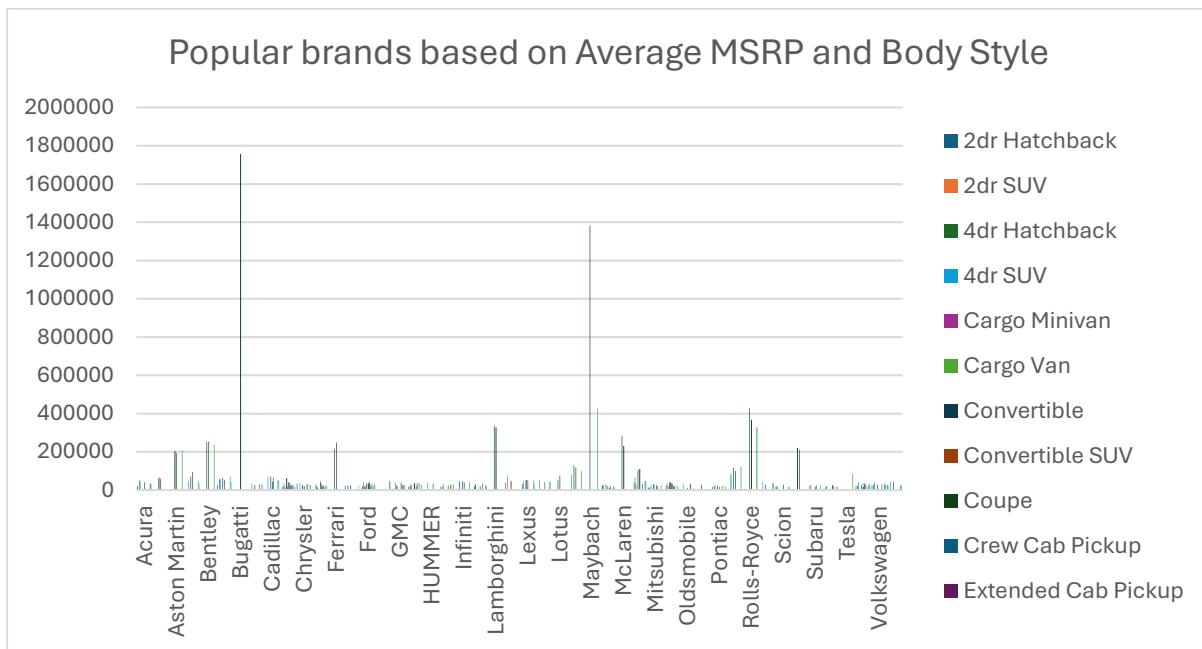
The contrast between the top and bottom brands underscores how pricing strategy reflects brand positioning and target audience. High-end luxury brands focus on exclusivity, while economy brands prioritize accessibility.

### **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

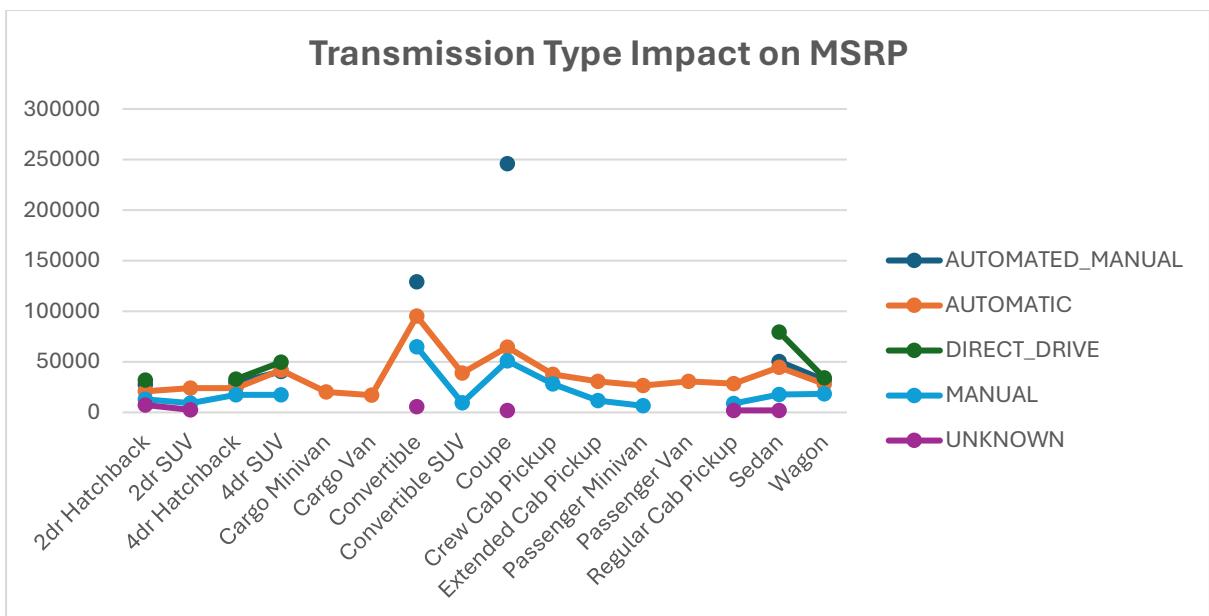
#### **Approach:**

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:**

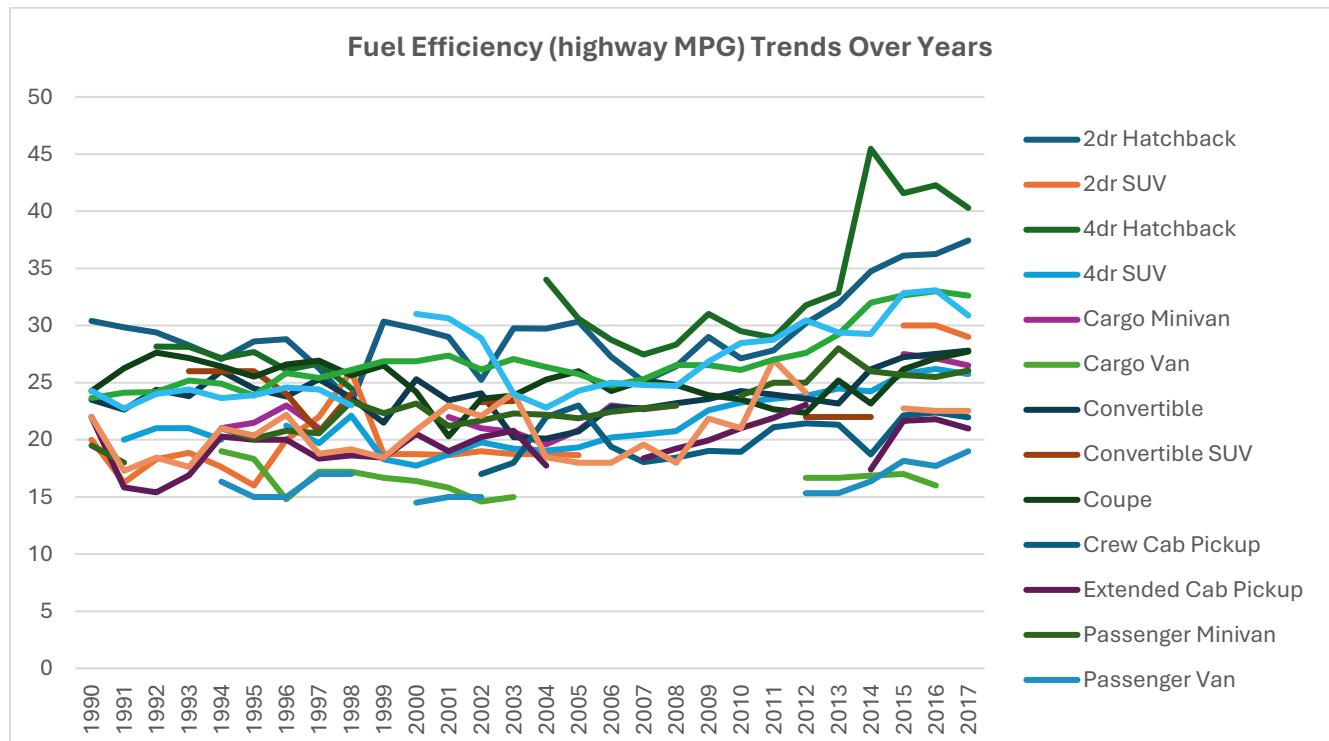
- 4dr Hatchbacks recorded the highest average MPG over the years, with values peaking above 45 MPG in later years like 2014 and beyond.
- 2dr Hatchbacks also consistently performed well, averaging 31.3 MPG over the entire period.

- **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.



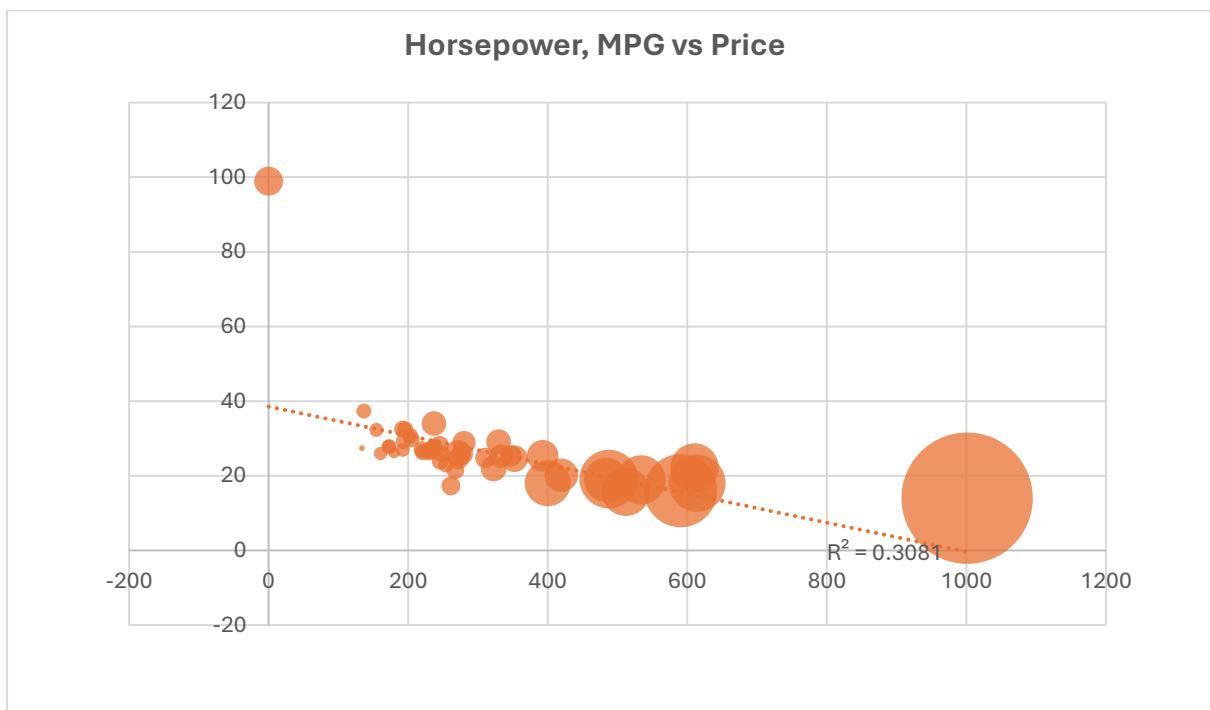
### Conclusion:

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](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   |
|----------------------------|---|
| <b>Microsoft Excel 365</b> | 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                 |
| <b>Grand Total</b> | <b>139.5321473</b>          |

### 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

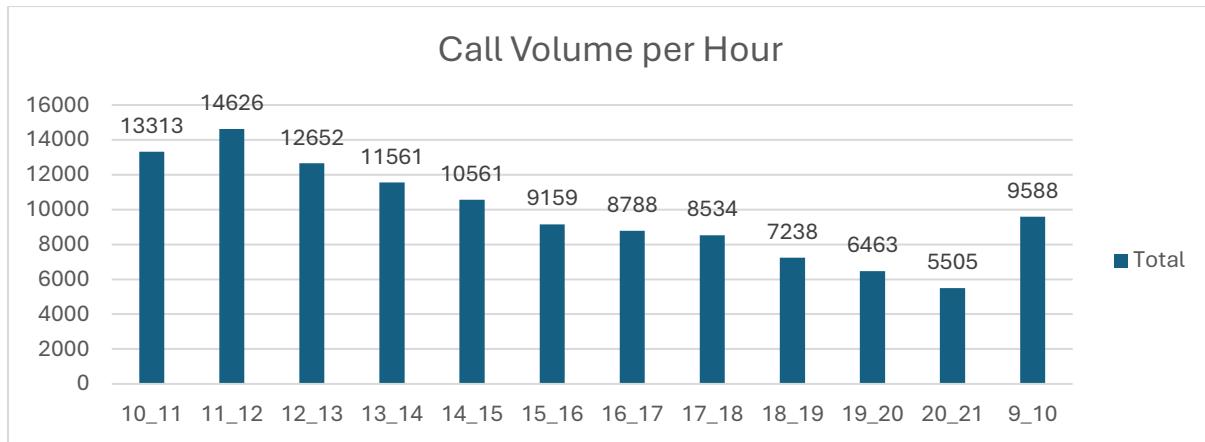
- The dataset was grouped into consistent hourly **Time Buckets** from **9:00 AM to 9:00 PM**, capturing the active window of customer service operations.
- Using **Pivot Tables**, the **number of calls** (approximated by count of entries in the **Call\_Seconds** column) was calculated for each time 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                      |
| <b>Grand Total</b> | <b>117988</b>             |

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:

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

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

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

$$\text{Adjusted Agent Hours} = \text{Agent Hours Required} \times (100 / 90)$$

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:

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

### 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, 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)

$$\text{Adjusted Agent Hours} = \text{Agent Hours} \times 1.42857$$

#### **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.

$$\text{Agents Needed} = [\text{Adjusted Agent Hours}]$$

### 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-zrG4XNmDoFNr0nxWLDBmf2NJEKgnBd/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true>



# IMDB MOVIE ANALYSIS

*PROJECT REPORT*

This project analyzes IMDB movie data using excel to uncover trends in genre, duration, language, director performance, and profitability. Key insights include popular genres, ideal movie length, top-performing directors, and a weak correlation between budget and gross revenue. The analysis provides data-driven insights for decision-making in the film industry.

Deepthy A  
Trainity Data Analytic Training

# Project Description

The IMDB Movie Analysis project was designed to explore the elements that contribute to a movie's success, both critically and commercially. By examining a dataset of over 5,000 movies, the goal was to uncover how various factors - such as genre, duration, language, director, and budget - relate to a film's IMDB rating and financial performance. This analysis aims to draw insights that can inform future decisions in filmmaking, marketing, and production strategy.

## Approach

Our approach began with cleaning and organizing the dataset to ensure accuracy and consistency. From there, we divided the analysis into five focused areas:

1. **Genre Analysis** - Evaluated how different genres perform on average in terms of IMDB scores.
2. **Movie Duration Analysis** - Explored how movie length affects viewer ratings, using descriptive statistics and visualizations.
3. **Language Analysis** - Assessed the frequency of languages used in movies and how language impacts IMDB ratings.
4. **Director Analysis** - Identified top-performing directors based on average scores and percentile rankings.
5. **Budget Analysis** - Investigated the financial side by analyzing the relationship between budgets, earnings, and profit margins.

Each section combined statistical functions to gain a clear, data-backed understanding of trends.

## Tech Stack Used

- **Microsoft Excel 365**  
Used extensively for data cleaning, statistical analysis (using formulas like AVERAGE, MEDIAN, STDEV, CORREL, etc.), pivot tables, and chart creation for visual insights.
- **Microsoft Word 365**  
Used for compiling, formatting, and documenting the report in a structured and professional format suitable for presentation and submission.

## Data Cleaning and Preprocessing

The initial dataset contained over 5,000 movie records with details such as genre, duration, language, director, budget, gross earnings, and IMDB scores. To ensure high data quality and reliable insights, several cleaning steps were carried out using **Power Query** in **Microsoft Excel 365**:

- **Removed blank rows and null entries:** Rows with empty or missing critical fields were filtered out to maintain consistency.
  - **Replaced inconsistent text entries:** Common issues such as inconsistent capitalization and extra spaces in genres and languages were resolved.
  - **Split multi-genre fields:** Movies with multiple genres were split into separate columns for more granular analysis.
  - **Changed data types:** Ensured numerical fields like "duration", "IMDB score", "budget", and "gross" were correctly formatted as numbers.
  - **Applied transformations:** Used Excel formulas (e.g., TRIM, LOWER, IFERROR) and Power Query transformations to clean and structure the data efficiently.

These preprocessing steps allowed for smooth execution of subsequent analysis tasks and improved the accuracy of insights generated.

File Home Transform Add Column View

Close & Load Refresh Advanced Editor Choose Columns Remove Rows Keep Rows Remove Rows Split Column Group By Use First Row as Headers Replace Values Data Type: Text Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Enter Data New Query

Close & Load Preview Manage Properties Column profiling based on top 1000 rows

Queries > = Table.TransformColumnTypes(#"Split Column by Delimiter",{{{"genres.1", type text}, {"genres.2", type text}, {"genres.3", type text}, {"genres.4", type text}, {"genres.5", type text}, {"genres.6", type text}, {"genres.7", type text}})

|    | genres.1  | genres.2  | genres.3 | genres.4 | genres.5 | genres.6 | genres.7 |
|----|-----------|-----------|----------|----------|----------|----------|----------|
| 1  | Action    | Adventure | Fantasy  | Sci-Fi   | null     | null     |          |
| 2  | Action    | Adventure | Fantasy  |          | null     | null     |          |
| 3  | Action    | Adventure | Thriller |          | null     | null     |          |
| 4  | Action    | Thriller  |          | null     | null     | null     |          |
| 5  | Action    | Adventure | Sci-Fi   |          | null     | null     |          |
| 6  | Action    | Adventure | Romance  |          | null     | null     |          |
| 7  | Adventure | Animation | Comedy   | Family   | Fantasy  | Musical  | Romance  |
| 8  | Action    | Adventure | Sci-Fi   |          | null     | null     |          |
| 9  | Adventure | Family    | Fantasy  | Mystery  |          | null     |          |
| 10 | Action    | Adventure | Sci-Fi   |          | null     | null     |          |
| 11 | Action    | Adventure | Sci-Fi   |          | null     | null     |          |
| 12 | Action    | Adventure |          | null     | null     | null     |          |
| 13 | Action    | Adventure | Fantasy  |          | null     | null     |          |
| 14 | Action    | Adventure | Western  |          | null     | null     |          |
| 15 | Action    | Adventure | Fantasy  | Sci-Fi   |          | null     |          |
| 16 | Action    | Adventure | Family   | Fantasy  |          | null     |          |
| 17 | Action    | Adventure | Sci-Fi   |          | null     | null     |          |
| 18 | Action    | Adventure | Fantasy  |          | null     | null     |          |
| 19 | Action    | Adventure | Comedy   | Family   | Fantasy  | Sci-Fi   |          |
| 20 | Adventure | Fantasy   |          | null     | null     | null     |          |
| 21 | Action    | Adventure | Fantasy  |          | null     | null     |          |
| 22 | Action    | Adventure | Drama    | History  |          | null     |          |

34 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 01:18 PM

Query Settings

Properties Name IMDb\_Movies All Properties

Applied Steps Source Promoted Headers Changed Type Removed Blank Rows Replaced Value Replaced Value1 Replaced Value2 Filtered Rows Replaced Value3 Cleared Text Split Column by Delimiter Changed Type1

# Movie Genre Analysis

**Objective:** To examine how different movie genres are associated with varying IMDB scores and assess which genres perform better overall.

## Approach:

- Extracted genre information using Power Query by splitting separated values into distinct genre columns.
- Transformed the data to consolidate genre frequency and calculate **median**, **mode**, and **average IMDB score** per genre using **Pivot Tables** and **Excel's statistical functions**.

| Genre       | Count | Median | Mode    |
|-------------|-------|--------|---------|
| Action      | 1147  | 6.3    | 6.1     |
| Adventure   | 921   | 6.6    | 6.7     |
| Fantasy     | 609   | 6.4    | 6.7     |
| Sci-Fi      | 615   | 6.4    | 6.7     |
| Thriller    | 1401  | 6.4    | 6.1     |
| Romance     | 1105  | 6.5    | 6.5     |
| Animation   | 241   | 6.7    | 6.7     |
| Comedy      | 1867  | 6.3    | 6.7     |
| Family      | 545   | 6.4    | 6.7     |
| Musical     | 132   | 6.7    | 7       |
| Mystery     | 498   | 6.6    | 6.6     |
| Western     | 97    | 6.8    | 6.5     |
| Drama       | 2582  | 6.9    | 7.2     |
| History     | 206   | 7.2    | 7.5     |
| Sport       | 182   | 6.8    | 7.2     |
| Crime       | 884   | 6.6    | 6.6     |
| Horror      | 561   | 5.9    | 6.2     |
| War         | 212   | 7.1    | 7.1     |
| Biography   | 293   | 7.2    | 7       |
| Music       | 214   | 6.6    | 6.5     |
| Documentary | 120   | 7.4    | 7.5     |
| Game-Show   | 1     | 2.9    | No Mode |
| Reality-TV  | 2     | 4.75   | No Mode |
| News        | 3     | 7.4    | No Mode |
| Short       | 5     | 6.5    | No Mode |
| Film-Noir   | 6     | 7.65   | No Mode |

| Row Labels  | Count of imdb_score | Var of imdb_score | StdDev of imdb_score | Min of imdb_score | Average of imdb_score | Max of imdb_score |
|-------------|---------------------|-------------------|----------------------|-------------------|-----------------------|-------------------|
| Action      | 1147                | 1.240020008       | 1.113561856          | 1.7               | 6.237314734           | 9                 |
| Adventure   | 921                 | 1.276122646       | 1.129655985          | 1.9               | 6.439087948           | 8.9               |
| Animation   | 241                 | 1.30269917        | 1.141358476          | 1.7               | 6.578423237           | 8.6               |
| Biography   | 293                 | 0.52202908        | 0.722515799          | 4.5               | 7.150170648           | 8.9               |
| Comedy      | 1867                | 1.1922221         | 1.091889234          | 1.7               | 6.194643814           | 9.5               |
| Crime       | 884                 | 1.050573118       | 1.024974691          | 2.4               | 6.564027149           | 9.3               |
| Documentary | 120                 | 1.125595938       | 1.060941063          | 1.6               | 7.180833333           | 8.7               |
| Drama       | 2582                | 0.90936991        | 0.953608887          | 2                 | 6.767505809           | 9.3               |
| Family      | 545                 | 1.446381409       | 1.202655981          | 1.7               | 6.245504587           | 8.7               |
| Fantasy     | 609                 | 1.349252063       | 1.161573099          | 1.7               | 6.307553366           | 8.9               |
| Film-Noir   | 6                   | 0.186666667       | 0.43204938           | 7.1               | 7.633333333           | 8.2               |
| History     | 206                 | 0.78747336        | 0.887396957          | 2                 | 7.088349515           | 8.9               |
| Horror      | 561                 | 1.266846893       | 1.125542933          | 2.2               | 5.853832442           | 8.7               |
| Music       | 214                 | 1.389659076       | 1.178838019          | 1.6               | 6.410280374           | 8.5               |
| Musical     | 132                 | 1.502384918       | 1.225718123          | 2.1               | 6.507575758           | 8.5               |
| Mystery     | 498                 | 1.188719506       | 1.09028414           | 2.2               | 6.48935743            | 8.6               |
| News        | 3                   | 0.263333333       | 0.513160144          | 7.1               | 7.533333333           | 8.1               |
| Reality-TV  | 2                   | 6.845             | 2.61629509           | 2.9               | 4.75                  | 6.6               |
| Romance     | 1105                | 0.993480507       | 0.996734923          | 2.1               | 6.450135747           | 8.6               |
| Sci-Fi      | 615                 | 1.458118588       | 1.207525813          | 1.9               | 6.277723577           | 8.8               |
| Short       | 5                   | 0.557             | 0.746324326          | 5.2               | 6.38                  | 7.1               |
| Sport       | 182                 | 1.214272661       | 1.101940407          | 2                 | 6.606043956           | 8.7               |
| Thriller    | 1401                | 1.099543683       | 1.048591285          | 2.2               | 6.319057816           | 9                 |
| War         | 212                 | 0.765911428       | 0.875163658          | 2.7               | 7.074056604           | 8.6               |
| Western     | 97                  | 1.086767612       | 1.042481468          | 3.8               | 6.689690722           | 8.9               |
| Grand Total | 14448               | 1.205277697       | 1.0978514            | 1.6               | 6.452249446           | 9.5               |

## Key Findings:

- The most frequent genres were **Drama (2582 movies)**, **Comedy (1867)**, and **Action (1147)**.
- Genres such as **Comedy**, **Crime**, and **Drama** displayed higher median IMDB scores.
- Less common genres like **Film-Noir**, though fewer in number, had the highest median (7.65) and average (7.63) IMDB scores.
- Popular genres like **Action** and **Comedy** had large sample sizes but relatively lower median scores (6.3 and 6.4).

## Insights:

- Niche or information-rich genres are rated more highly by audiences, possibly due to deeper storytelling or educational value.
- Commercial genres like **Comedy** and **Action** are more widespread but tend to cluster around average ratings.

# Movie Duration Analysis

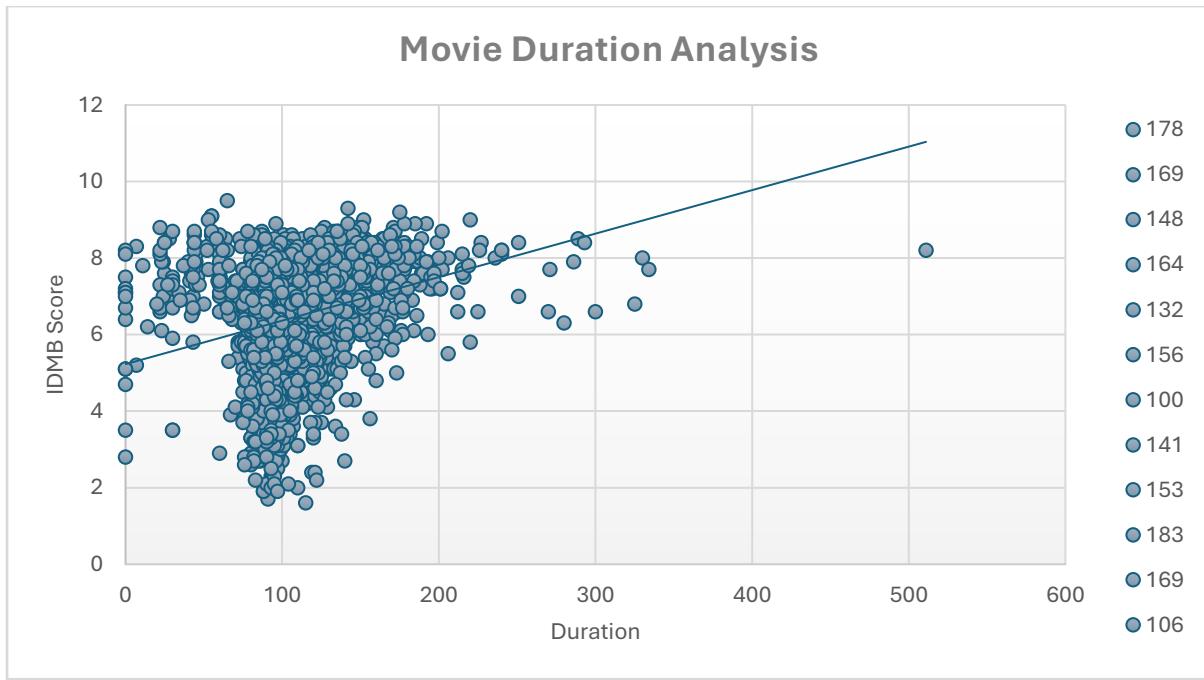
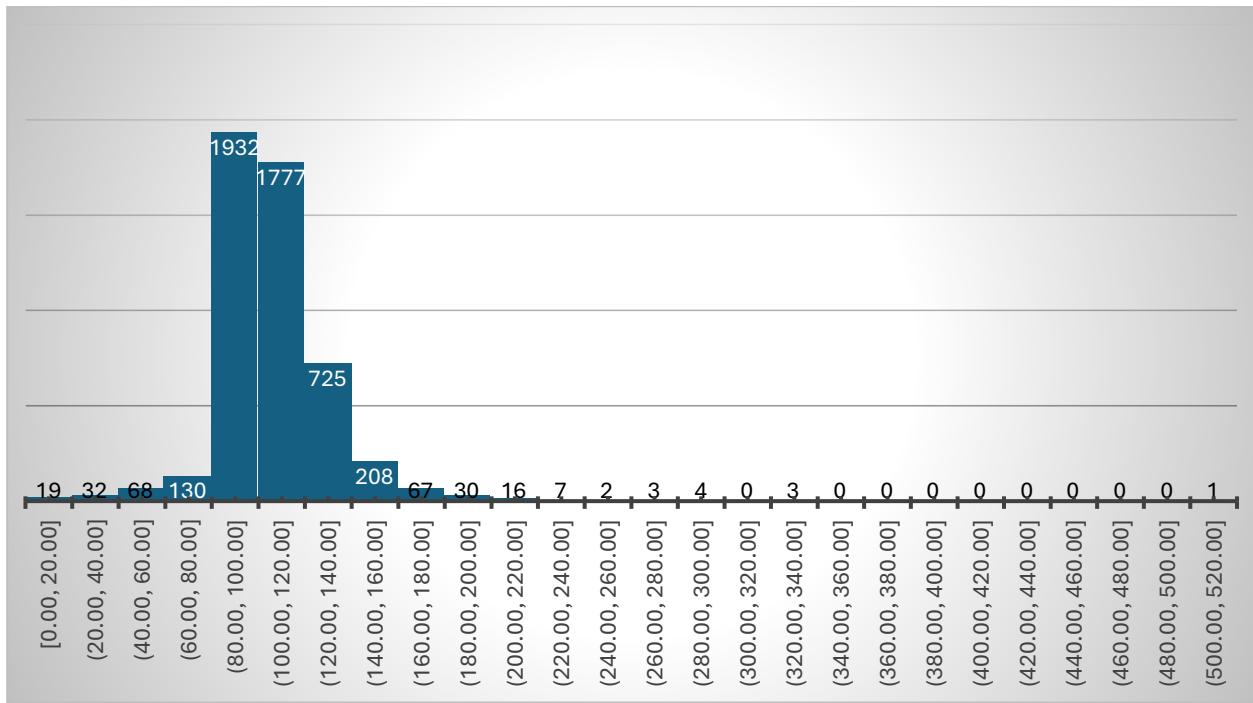
## Objective:

To explore the relationship between a movie's **duration** and its **IMDB score**, identifying any trends or outliers.

## Approach:

- Used a **scatter plot with a trendline** to visualize the correlation between Duration and IMDB Score.
- Calculated **Mean, Median, and Standard Deviation** of durations for each genre to understand how movie lengths vary by category.

| Genre      | Mean     | Median | Stdev    |
|------------|----------|--------|----------|
| Action     | 110.8387 | 107    | 1.113562 |
| Adventure  | 112.0456 | 107    | 1.129656 |
| Fantasy    | 103.087  | 100    | 1.161573 |
| Sci-Fi     | 107.187  | 104    | 1.207526 |
| Thriller   | 107.935  | 106    | 1.048591 |
| Romance    | 108.867  | 105    | 0.996735 |
| Animation  | 88.54357 | 90     | 1.141358 |
| Comedy     | 98.91912 | 98     | 1.091889 |
| Family     | 97.00183 | 95     | 1.202656 |
| Musical    | 109.5076 | 103    | 1.225718 |
| Mystery    | 106.51   | 105    | 1.090284 |
| Western    | 120.1443 | 110    | 1.042481 |
| Drama      | 112.6979 | 109    | 0.953609 |
| History    | 136.5485 | 131    | 0.887397 |
| Sport      | 110.2253 | 108.5  | 1.10194  |
| Crime      | 108.336  | 105    | 1.024975 |
| Horror     | 97.89483 | 96     | 1.125543 |
| War        | 130.9245 | 124.5  | 0.875164 |
| Biography  | 125.0239 | 121    | 0.722516 |
| Music      | 108.3084 | 107    | 1.178838 |
| Document   | 93.70833 | 92.5   | 1.060941 |
| Game-Shc   | 60       | 60     | no stdev |
| Reality-TV | 51.5     | 51.5   | 2.616295 |
| News       | 98.66667 | 105    | 0.51316  |
| Short      | 26.4     | 34     | 0.746324 |
| Film-Noir  | 99.5     | 93.5   | 0.432049 |



#### Interpretation:

- **Longer movies** (especially History, Biography, War, and Drama) generally correlate with **higher ratings**.
- **Short formats** like Game Shows, Reality-TV, and Shorts have **lower durations and lower average ratings**, suggesting audience preferences might lean toward longer, narrative-rich content.

- **Standard deviations are low**, showing that most genres have consistent duration patterns, with **Reality-TV** being an exception (highest variation).

#### **Key Visual Insight:**

The scatter plot reveals a **slight positive correlation** between movie duration and IMDB score. As the duration increases, IMDB scores also tend to rise marginally. However, the spread is wide, indicating that **duration alone is not a strong predictor** of a high rating.

# Language Analysis

## Objective:

To determine whether **movie language** influences **IMDB scores**, by analyzing average, median ratings and standard deviation per language.

## Approach:

- Aggregated data for each language including **Mean**, **Median**, and **Standard Deviation** of IMDB ratings.
- Highlighted high-performing and low-performing languages.
- Created a **conditional formatting heatmap** for quick visual insights.

| Language   | Count | Mean     | Median | Stdev       |
|------------|-------|----------|--------|-------------|
| English    | 4692  | 6.400554 | 6.5    | 1.120447972 |
| Japanese   | 18    | 7.394444 | 7.6    | 0.990823913 |
| French     | 73    | 7.038356 | 7.2    | 0.726985812 |
| Mandarin   | 25    | 6.776    | 7      | 1.061555463 |
| Aboriginal | 2     | 6.95     | 6.95   | 0.777817459 |
| Spanish    | 40    | 6.9375   | 7.15   | 0.855056603 |
| Filipino   | 1     | 6.7      | 6.7    |             |
| Hindi      | 28    | 6.632143 | 6.95   | 1.398955582 |
| Russian    | 11    | 6.363636 | 6.5    | 1.383671007 |
| Maya       | 1     | 7.8      | 7.8    |             |
| Kazakh     | 1     | 6        | 6      |             |
| Telugu     | 1     | 8.4      | 8.4    |             |
| Cantonese  | 11    | 6.954545 | 7.2    | 0.704788814 |
| Icelandic  | 2     | 7.55     | 7.55   | 0.919238816 |
| German     | 19    | 7.342105 | 7.6    | 0.954123093 |
| Aramaic    | 1     | 7.1      | 7.1    |             |
| Italian    | 11    | 7.227273 | 7.3    | 1.244259546 |
| Dutch      | 4     | 7.425    | 7.45   | 0.434932945 |
| Dari       | 2     | 7.5      | 7.5    | 0.141421356 |
| Hebrew     | 5     | 7.58     | 7.6    | 0.334664011 |
| Chinese    | 3     | 5.666667 | 5.7    | 0.550757055 |
| Mongolian  | 1     | 7.3      | 7.3    |             |
| Swedish    | 5     | 7.44     | 7.6    | 0.756967635 |
| Korean     | 8     | 7.3875   | 7.5    | 0.825378701 |
| Thai       | 3     | 6.633333 | 6.6    | 0.450924975 |
| Polish     | 4     | 8.25     | 8.25   | 0.981495458 |
| Bosnian    | 1     | 4.3      | 4.3    |             |
| None       | 2     | 7.95     | 7.95   | 0.777817459 |
| Hungarian  | 1     | 7.1      | 7.1    |             |
| Portuguese | 8     | 7.4875   | 7.7    | 0.883883476 |

|            |   |       |      |             |
|------------|---|-------|------|-------------|
| Unknown    | 6 | 6.8   | 6.95 | 1.331164903 |
| Danish     | 5 | 7.5   | 8.1  | 1.077032961 |
| Arabic     | 5 | 7.38  | 7.4  | 0.884307639 |
| Norwegian  | 4 | 7.15  | 7.3  | 0.574456265 |
| Czech      | 1 | 7.4   | 7.4  |             |
| Kannada    | 1 | 7.1   | 7.1  |             |
| Zulu       | 2 | 7.1   | 7.1  | 0.282842712 |
| Punjabi    | 1 | 6.6   | 6.6  |             |
| Tamil      | 1 | 5.1   | 5.1  |             |
| Dzongkha   | 1 | 7.5   | 7.5  |             |
| Vietnamese | 1 | 7.4   | 7.4  |             |
| Indonesian | 2 | 7.9   | 7.9  | 0.424264069 |
| Urdu       | 1 | 7     | 7    |             |
| Romanian   | 2 | 7.2   | 7.2  | 0.989949494 |
| Persian    | 4 | 7.575 | 7.95 | 1.203813385 |
| Slovenian  | 1 | 6.4   | 6.4  |             |
| Greek      | 1 | 7.3   | 7.3  |             |
| Swahili    | 1 | 7.4   | 7.4  |             |

### Key Insights:

#### 1. Highest Rated Languages (Mean Rating > 7.5)

- **Telugu (8.4)** - Highest rated, though based on a single entry.
- **Polish (8.25)** - Consistently high rating, stronger due to multiple entries.
- **Maya (7.8), Indonesian (7.9), and Hebrew(7.58)** also show strong average ratings with low standard deviations, indicating consistency.

#### 2. Popular Languages with High Volume:

- **English** (4,692 movies): Mean score **6.4**, Median **6.5**, with a wider spread (Stdev: 1.12).
  - Indicates **broad diversity** in English-language films ranging from low to high-rated.
- **French** (73): High mean rating (**7.03**) and low deviation (**0.72**) - shows reliable performance.
- **Japanese** and **German** are also strong performers with means above 7.3.

#### 3. Languages with Lower Scores:

- **Chinese** (Mean: 5.67), **Bosnian** (4.3), and **Tamil** (5.1) are among the **lowest rated**.
- These may represent niche or limited-audience productions, or limited data points.

4. **Languages with Consistent Ratings (Low Std Dev):**

- **Dari (0.14), Slovenian (N/A), Dutch (0.43), and Indonesian (0.42)** show **tight clustering**, suggesting **strong consistency** in ratings.

# Director Analysis

## Objective:

To identify the top-performing directors based on average IMDb movie ratings and understand the characteristics of high-rated direction.

## Approach:

- Calculated the 95th percentile threshold using the formula:  
`=PERCENTILE.EXC(B:B, 0.95)`

| =PERCENTILE.EXC(B:B, 0.95) |        |   |   |   |        |
|----------------------------|--------|---|---|---|--------|
|                            | B      | C | D | E | F      |
| Mean                       | Top 5% |   |   |   |        |
| 7.914286                   | Top 5% |   |   |   |        |
| 6.985714                   |        |   |   |   | 7.7925 |
| 7.5                        |        |   |   |   |        |

- Directors with a mean rating equal to or above **7.79** were classified as **Top 5%**.
- Tagged and extracted data for directors meeting the criteria.

| Director          | Mean     | Top 5% |
|-------------------|----------|--------|
| James Cameron     | 7.914286 | Top 5% |
| Christopher Nolan | 8.425    | Top 5% |
| Nathan Greno      | 7.8      | Top 5% |
| Joss Whedon       | 7.925    | Top 5% |
| Lee Unkrich       | 8.3      | Top 5% |
| Pete Docter       | 8.233333 | Top 5% |
| Don Hall          | 7.9      | Top 5% |
| Rich Moore        | 7.8      | Top 5% |
| Hideaki Anno      | 8.2      | Top 5% |
| Alfonso Cuarón    | 7.8      | Top 5% |
| Quentin Tarantino | 8.2      | Top 5% |
| Jacques Perrin    | 7.9      | Top 5% |
| Frank Darabont    | 7.975    | Top 5% |
| Stanley Kubrick   | 8        | Top 5% |
| Tim Miller        | 8.1      | Top 5% |
| Milos Forman      | 8.133333 | Top 5% |
| Deepa Mehta       | 7.8      | Top 5% |
| Andrei Tarkovsky  | 8.1      | Top 5% |
| Denis Villeneuve  | 7.966667 | Top 5% |
| S.S. Rajamouli    | 8.4      | Top 5% |
| Moustapha Akkad   | 8.4      | Top 5% |

|                         |          |        |
|-------------------------|----------|--------|
| Tony Kaye               | 8.033333 | Top 5% |
| Hayao Miyazaki          | 8.225    | Top 5% |
| Richard Marquand        | 8.4      | Top 5% |
| Sergio Leone            | 8.475    | Top 5% |
| David Lean              | 8        | Top 5% |
| Bernardo Bertolucci     | 7.95     | Top 5% |
| Giuseppe Tornatore      | 7.8      | Top 5% |
| Christian Carion        | 7.8      | Top 5% |
| Tom McCarthy            | 7.9      | Top 5% |
| James Schamus           | 7.8      | Top 5% |
| Terry George            | 8.1      | Top 5% |
| George Cukor            | 7.9      | Top 5% |
| Michel Hazanavicius     | 8        | Top 5% |
| Morten Tyldum           | 7.85     | Top 5% |
| Catherine Owens         | 8.4      | Top 5% |
| Fritz Lang              | 8.3      | Top 5% |
| Andrey Zvyagintsev      | 8        | Top 5% |
| John Blanchard          | 9.5      | Top 5% |
| Lenny Abrahamson        | 8.3      | Top 5% |
| Stephen Chbosky         | 8        | Top 5% |
| John Cromwell           | 7.8      | Top 5% |
| Je-kyu Kang             | 8.1      | Top 5% |
| Stéphane Aubier         | 7.9      | Top 5% |
| Josh Boone              | 7.8      | Top 5% |
| Akira Kurosawa          | 8.1      | Top 5% |
| Ken Annakin             | 7.8      | Top 5% |
| Stanley Kramer          | 7.95     | Top 5% |
| Dan Gilroy              | 7.9      | Top 5% |
| Jonathan Dayton         | 7.9      | Top 5% |
| Justin Tipping          | 7.8      | Top 5% |
| Vincent Paronnaud       | 8        | Top 5% |
| George Roy Hill         | 8.2      | Top 5% |
| Robert Stevenson        | 7.8      | Top 5% |
| Christophe Barratier    | 7.9      | Top 5% |
| Rakeysh Omprakash Mehra | 8.4      | Top 5% |
| Mike Mayhall            | 8.6      | Top 5% |
| Shona Auerbach          | 7.8      | Top 5% |
| Mike van Diem           | 7.8      | Top 5% |
| Lukas Moodysson         | 7.9      | Top 5% |
| Raja Menon              | 8.5      | Top 5% |
| Billy Wilder            | 7.975    | Top 5% |
| John Sturges            | 8.3      | Top 5% |
| Ron Fricke              | 8.5      | Top 5% |
| Howard Hughes           | 7.8      | Top 5% |
| Jim Abrahams            | 7.8      | Top 5% |
| Richard Brooks          | 7.8      | Top 5% |
| Jay Oliva               | 8.4      | Top 5% |
| Charles Ferguson        | 7.866667 | Top 5% |

|                       |          |        |
|-----------------------|----------|--------|
| Damien Chazelle       | 8.5      | Top 5% |
| Frank Capra           | 8.06     | Top 5% |
| Howard Hawks          | 7.8      | Top 5% |
| Mel Stuart            | 7.8      | Top 5% |
| Rajkumar Hirani       | 8.2      | Top 5% |
| Stacy Peralta         | 7.8      | Top 5% |
| Stanley Donen         | 8.3      | Top 5% |
| William Wyler         | 8.1      | Top 5% |
| Robert Mulligan       | 8.4      | Top 5% |
| Robert Rossen         | 8        | Top 5% |
| Sylvain Chomet        | 7.8      | Top 5% |
| Elia Kazan            | 7.866667 | Top 5% |
| Ari Folman            | 8        | Top 5% |
| David Sington         | 8.1      | Top 5% |
| Ralph Ziman           | 7.8      | Top 5% |
| Jehane Noujaim        | 8.1      | Top 5% |
| Mitchell Altieri      | 8.7      | Top 5% |
| Ritesh Batra          | 7.8      | Top 5% |
| Fabián Bielinsky      | 7.9      | Top 5% |
| Charles Chaplin       | 8.6      | Top 5% |
| Frank Lotito          | 8.2      | Top 5% |
| Sadyk Sher-Niyaz      | 8.7      | Top 5% |
| Anna Muylaert         | 7.9      | Top 5% |
| Robert Kenner         | 7.9      | Top 5% |
| Joshua Oppenheimer    | 8.2      | Top 5% |
| Cristian Mungiu       | 7.9      | Top 5% |
| Asghar Farhadi        | 8.4      | Top 5% |
| Georg Wilhelm Pabst   | 8        | Top 5% |
| Mark Sandrich         | 7.8      | Top 5% |
| Michael Wadleigh      | 8.1      | Top 5% |
| Lance McDaniel        | 8        | Top 5% |
| Marius A. Markevicius | 8.4      | Top 5% |
| Justin Paul Miller    | 8.3      | Top 5% |
| Joe Kenemore          | 8.2      | Top 5% |
| Ingmar Bergman        | 8.2      | Top 5% |
| D.W. Griffith         | 8        | Top 5% |
| Henry Alex Rubin      | 7.8      | Top 5% |
| Lauren Lazin          | 8        | Top 5% |
| Michael Roemer        | 8.1      | Top 5% |
| Sam Martin            | 7.8      | Top 5% |
| Kristin Rizzo         | 8        | Top 5% |
| Carl Theodor Dreyer   | 8.1      | Top 5% |
| Sharon Greytak        | 8.1      | Top 5% |
| Majid Majidi          | 8.5      | Top 5% |
| Cary Bell             | 8.7      | Top 5% |
| Bill Melendez         | 8.4      | Top 5% |
| Amal Al-Agroobi       | 8.2      | Top 5% |
| Sut Jhally            | 8.3      | Top 5% |

|                 |     |        |
|-----------------|-----|--------|
| Dave Carroll    | 7.9 | Top 5% |
| Anthony Vallone | 7.8 | Top 5% |

### Key Insights:

- The **Top 5% threshold** was determined to be **7.79**.
- John Blanchard** holds the highest average rating of **9.5**.
- Some other directors are:

| Director         | Mean |
|------------------|------|
| John Blanchard   | 9.5  |
| Mike Mayhall     | 8.6  |
| Raja Menon       | 8.5  |
| Ron Fricke       | 8.5  |
| Damien Chazelle  | 8.5  |
| Mitchell Altieri | 8.7  |
| Charles Chaplin  | 8.6  |
| Sadyk Sher-Niyaz | 8.7  |
| Majid Majidi     | 8.5  |
| Cary Bell        | 8.7  |

- A total of **118 directors** fell into the Top 5% category.
- Many top directors have strong critical acclaim, cross-genre consistency, or a history of award-winning work.

# Budget Analysis

## Objective:

To explore the relationship between a movie's budget and its gross revenue and identify the most and least profitable titles.

## Approach:

- Calculated the **profit margin** using the formula:  
 $= \text{gross} - \text{budget}$
- Used the CORREL() function to measure the linear relationship between **budget** and **gross**.

```
=CORREL(IMDB_Movies[budget],IMDB_Movies[gross])
```

0.102166

- Identified the **highest** profit margin movie.

```
=INDEX(IMDB_Movies[movie_title],MATCH(MAX(D:D), D:D, 0))
```

| E   | F      | G | H         | I | J |
|-----|--------|---|-----------|---|---|
| 847 |        |   |           |   |   |
| 152 |        |   |           |   |   |
| 825 | Max    |   | 523505847 |   |   |
| 642 |        |   |           |   |   |
| 321 |        |   |           |   |   |
| 303 | Avatar |   |           |   |   |
| 738 |        |   |           |   |   |
| 599 |        |   |           |   |   |
| 980 |        |   |           |   |   |
| 062 |        |   |           |   |   |

- Found top 5 most profitable movies.

```
G10 : X ✓ fx =INDEX(IMDB_Movies[movie_title],MATCH(H10:H14,D2:D4996, 0))
```

| G                                  | H         | I | J | Formula Bar | M | N | O |
|------------------------------------|-----------|---|---|-------------|---|---|---|
| Avatar                             | 523505847 |   |   |             |   |   |   |
| Titanic                            | 458672302 |   |   |             |   |   |   |
| Jurassic World                     | 502177271 |   |   |             |   |   |   |
| Star Wars: Episode IV - A New Hope | 449935665 |   |   |             |   |   |   |
| E.T. the Extra-Terrestrial         | 424449459 |   |   |             |   |   |   |

### **Key Insights:**

- The **correlation coefficient** between budget and gross is **0.10**, suggesting a **very weak positive correlation**.
- This indicates that a **higher budget does not strongly guarantee higher revenue**.
- The **most profitable movie** is “**Avatar**”.
- Top 5 profitable movies are:

|   |
|---|
| <b>Avatar</b>                             |
| <b>Titanic</b>                            |
| <b>Jurassic World</b>                     |
| <b>Star Wars: Episode IV - A New Hope</b> |
| <b>E.T. the Extra-Terrestrial</b>         |

- Profitability varies widely, indicating that **factors beyond budget** (like content quality, marketing, or star power) play a major role in box office success.

## Conclusion

This analysis of IMDB movie data provided valuable insights into the elements that contribute to a movie's success. We identified the most popular genres, ideal movie durations, top-performing languages and directors, and examined the financial performance of films. While budget and gross revenue showed only a weak correlation, the study highlights that factors like genre and director can significantly influence a movie's reception. These findings can support better planning and decision-making for filmmakers and studios aiming to optimize both audience engagement and profitability.

Drive Link:

<https://docs.google.com/spreadsheets/d/1gslZabkwBWDtmNqlN4laboWrdxkfe2JT/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true>



# Hiring Process Analytics

Data Analysis Project

Deepthy A

Trainity Data Analytic Trainee

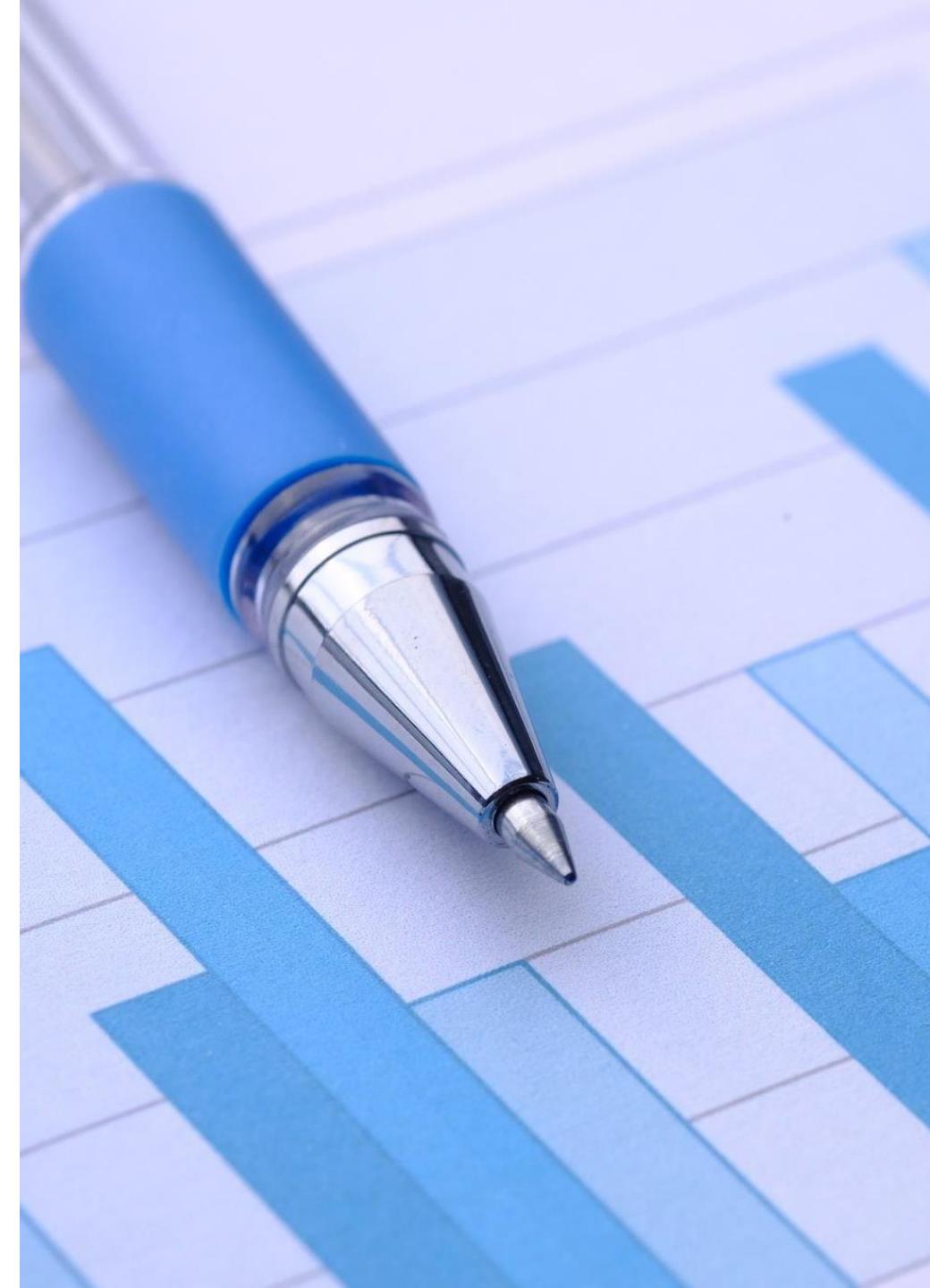
# Project Description

**Objective:** To analyze the hiring process data of a multinational company, uncover trends in hiring, salary, departmental distribution, and provide insights that can improve HR practices.

**Dataset:** Historical hiring records, including gender, salary, job titles, departments, etc.

# Introduction

- The Hiring Process Analytics project aims to analyze recruitment-related data to identify key insights and trends in hiring practices.
- This includes evaluating salary distributions, gender-based hiring patterns, average offered salaries, departmental hiring counts, and position tier distributions.
- Data cleaning and preprocessing were performed in Power Query, followed by analysis using pivot tables and visualizations in Excel.



# Tech-Stack Used

## Microsoft Excel 365

- Used for data cleaning, statistical computation, visualization (pivot tables, charts, etc.)

## Microsoft Powerpoint 365

- Used for creating report.

# Data Analytics Tasks



**Hiring Analysis:** The hiring process involves bringing new individuals into the organization for various roles.

- Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company?

**Salary Analysis:** The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.

- Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.

# Data Analytics Tasks



**Salary Distribution:** Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.

- Your Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

**Departmental Analysis:** Visualizing data through charts and plots is a crucial part of data analysis.

- Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

**Position Tier Analysis:** Different positions within a company often have different tiers or levels.

- Your Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

# Data Cleaning and Preprocessing

|             |           |
|-------------|-----------|
| Q1          | 25584.5   |
| Q3          | 74438     |
| IQR         | 48853.5   |
| Lower Bound | -47695.75 |
| Upper Bound | 147718.25 |

All data cleaning steps were conducted using Power Query:

- Removed duplicate rows.
- Handled null values by replacing or filtering them appropriately.
- Standardized column names and data types.
- Created conditional columns for status-based filtering.
- Applied outlier detection for salary using the IQR method:
  - Calculated Q1 and Q3.
  - Computed  $IQR = Q3 - Q1$ .
  - Filtered out salaries below  $Q1 - 1.5IQR$  or above  $Q3 + 1.5IQR$ .

## Hiringdata - Power Query Editor

**File** Home Transform Add Column View

Close & Load Refresh Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Sort Data Type: Whole Number Use First Row as Headers Merge Queries Append Queries Manage Parameters Data source settings New Source Recent Sources Enter Data Close Query Manage Columns Reduce Rows Group By Split Column Replace Values Combine Files Parameters Data Sources New Query

**Queries**

= Table.SelectRows(#"Sorted Rows", each [Offered Salary] <= 147784.5 and [Offered Salary] >= 1000)

|    | application_id | Interview Date | Interview Time | Status   | event_name        | Department            | Post Name |
|----|----------------|----------------|----------------|----------|-------------------|-----------------------|-----------|
| 1  | 366299         | 11-08-2014     | 02:58:13       | Rejected | Male              | Marketing Department  | c9        |
| 2  | 464899         | 14-05-2014     | 11:57:18       | Hired    | Female            | General Management    | i7        |
| 3  | 784107         | 19-05-2014     | 10:39:51       | Hired    | Male              | Operations Department | c9        |
| 4  | 699493         | 15-07-2014     | 11:34:41       | Rejected | Male              | Service Department    | c8        |
| 5  | 365930         | 01-07-2014     | 12:18:15       | Hired    | Female            | Finance Department    | c5        |
| 6  | 926938         | 01-05-2014     | 07:50:27       | Hired    | Female            | Service Department    | c8        |
| 7  | 766041         | 01-08-2014     | 09:32:19       | Rejected | Male              | Service Department    | i6        |
| 8  | 185368         | 27-06-2014     | 18:42:07       | Rejected | Male              | Service Department    | c5        |
| 9  | 300037         | 10-06-2014     | 09:32:01       | Hired    | Female            | Service Department    | b9        |
| 10 | 296052         | 29-05-2014     | 14:43:35       | Hired    | Female            | Finance Department    | c5        |
| 11 | 535514         | 22-07-2014     | 11:39:48       | Hired    | Male              | Operations Department | c5        |
| 12 | 347068         | 08-05-2014     | 15:03:18       | Hired    | Male              | Operations Department | c9        |
| 13 | 861850         | 11-06-2014     | 13:01:28       | Hired    | Male              | Operations Department | b9        |
| 14 | 169266         | 20-08-2014     | 18:51:07       | Rejected | Female            | Service Department    | c9        |
| 15 | 990318         | 24-05-2014     | 07:09:33       | Hired    | Female            | Production Department | i6        |
| 16 | 45162          | 19-05-2014     | 00:05:18       | Hired    | Don't want to say | Operations Department | i4        |
| 17 | 239190         | 20-06-2014     | 12:59:14       | Rejected | Male              | Operations Department | c9        |
| 18 | 47896          | 21-08-2014     | 17:46:33       | Rejected | Male              | Service Department    | b9        |
| 19 | 504811         | 14-05-2014     | 15:17:01       | Hired    | Male              | Purchase Department   | i6        |
| 20 | 772086         | 16-07-2014     | 18:21:18       | Hired    | Female            | Operations Department | c9        |
| 21 | 943922         | 24-06-2014     | 17:47:16       | Hired    | Male              | Operations Department | i7        |
| 22 | 490577         | 26-08-2014     | 18:47:33       | Hired    | Male              | Service Department    | i5        |
| 23 |                |                |                |          |                   |                       |           |

8 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

**Query Settings**

**PROPERTIES**

- Name: Hiringdata

All Properties

**APPLIED STEPS**

- Source
- Changed Type
- Replaced Value
- Removed Duplicates
- Split Column by Delimiter
- Changed Type1
- Renamed Columns
- Changed Type2
- Removed Blank Rows
- Replaced Value1
- Sorted Rows
- Filtered Rows

PREVIEW DOWNLOADED AT 08:53 PM

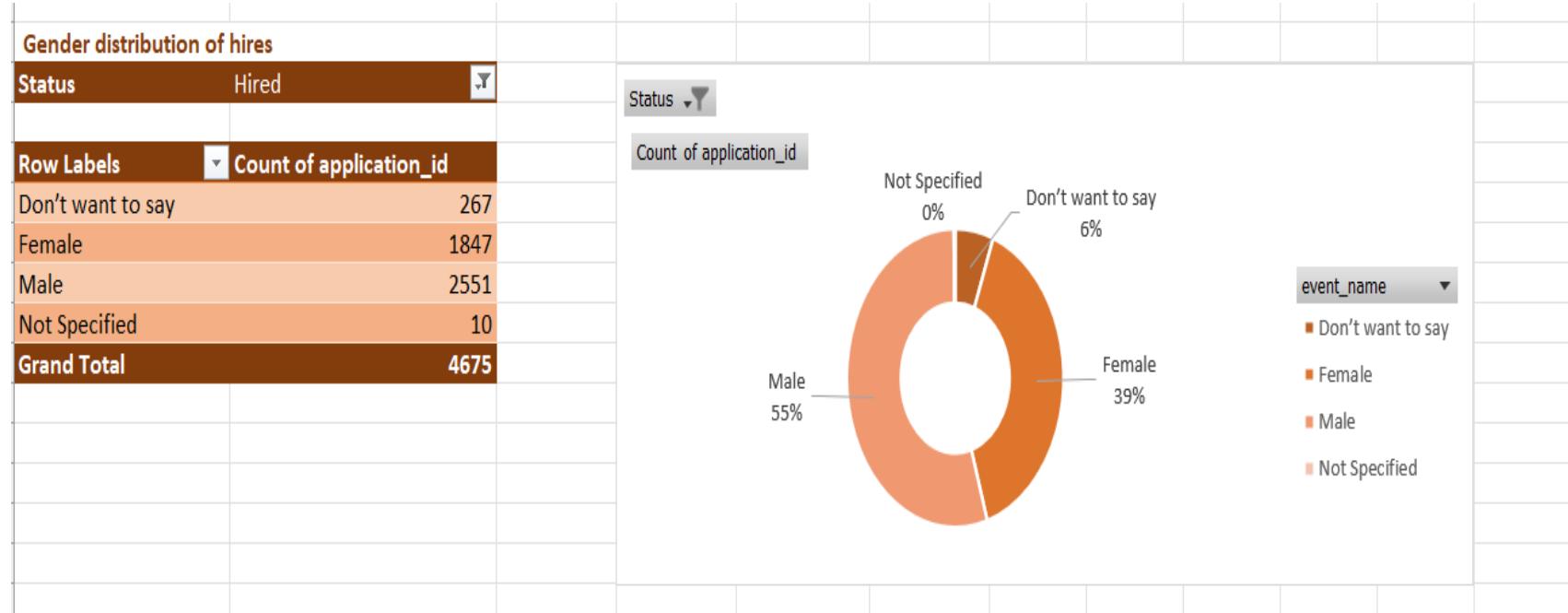
# Descriptive Statistics

These statistics provide an overview of the salary range and its variability in the dataset.

| Data Summary       |             |
|--------------------|-------------|
| Count              | 7135        |
| Mean               | 49918.73693 |
| Mode               | 20666       |
| Median             | 49636       |
| Minimum            | 1007        |
| Maximum            | 99967       |
| Standard Deviation | 28339.90332 |

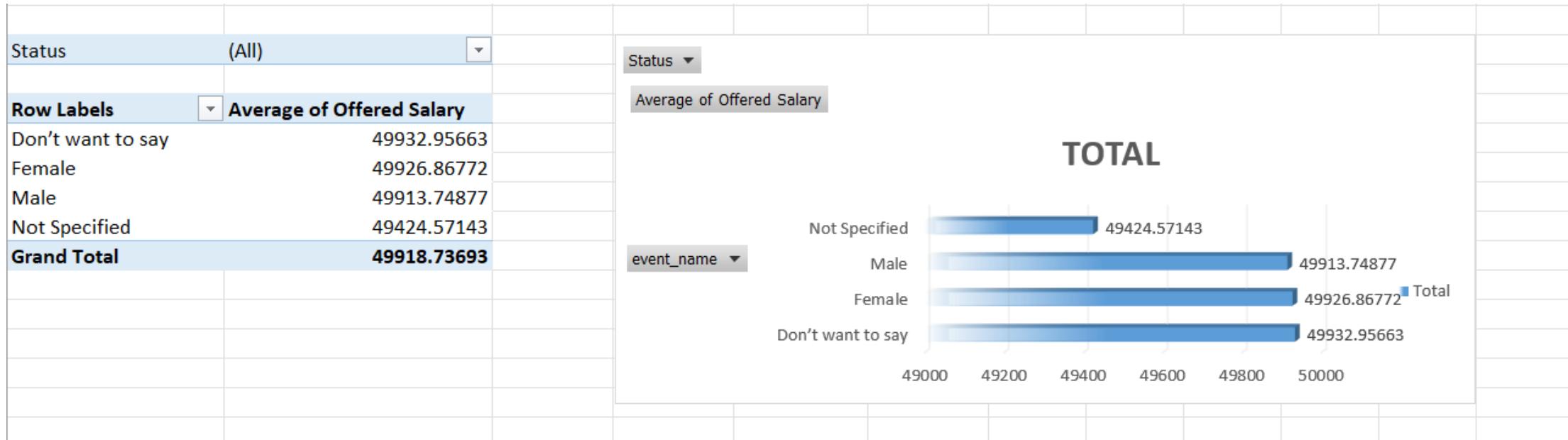
# Gender Distribution of Hires

- Male: 55%
- Female: 39%
- Others/Not specified: 6%
- The organization has a male-dominant hiring pattern. The share of female hires is significant but lags behind.



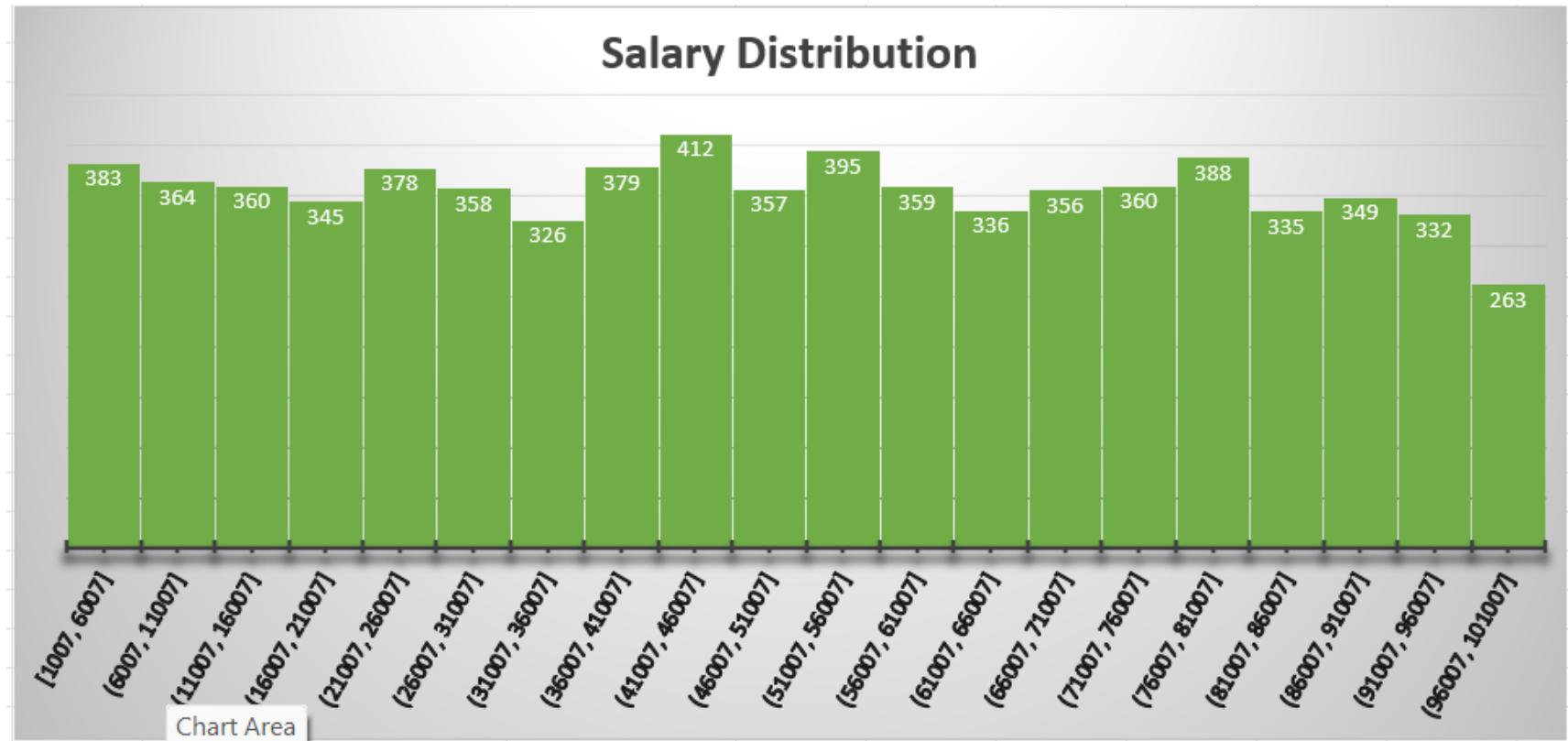
# Average Salary by Gender

- Average offered salary is quite consistent across gender identities.
- Slight variation (~₹20–₹30 difference), which is negligible statistically.
- Suggests no obvious gender-based bias in salary offers, which is a positive sign for equitable pay practices.



# Salary Distribution

- Most of the offered salaries fall within the ₹40,007–₹60,007 range, with peaks around ₹46,007–₹50,007 and ₹50,007–₹56,007.
- The salary distribution is relatively even with a slight peak in the mid-ranges. There's a steep drop in the ₹96,007–₹1,00,007 bracket, indicating fewer high-end offers.

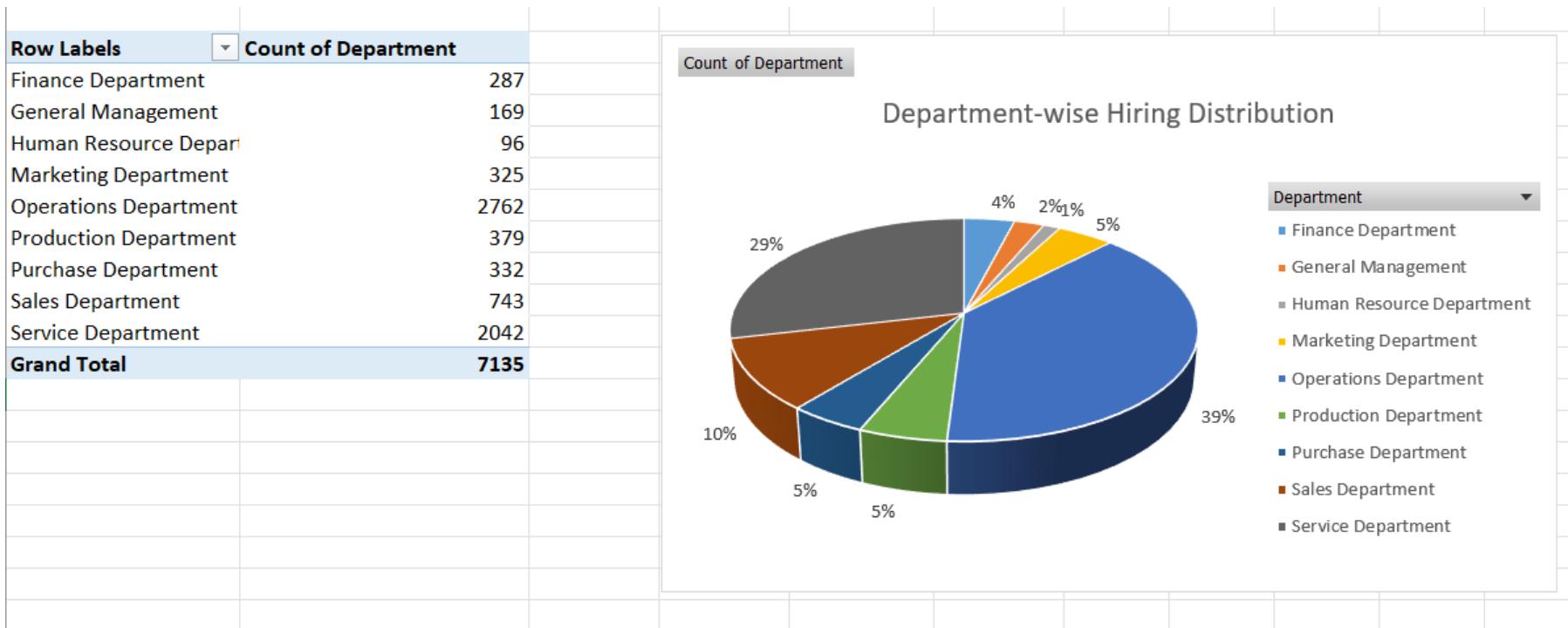


# Department-Wise Hiring

Top Departments:

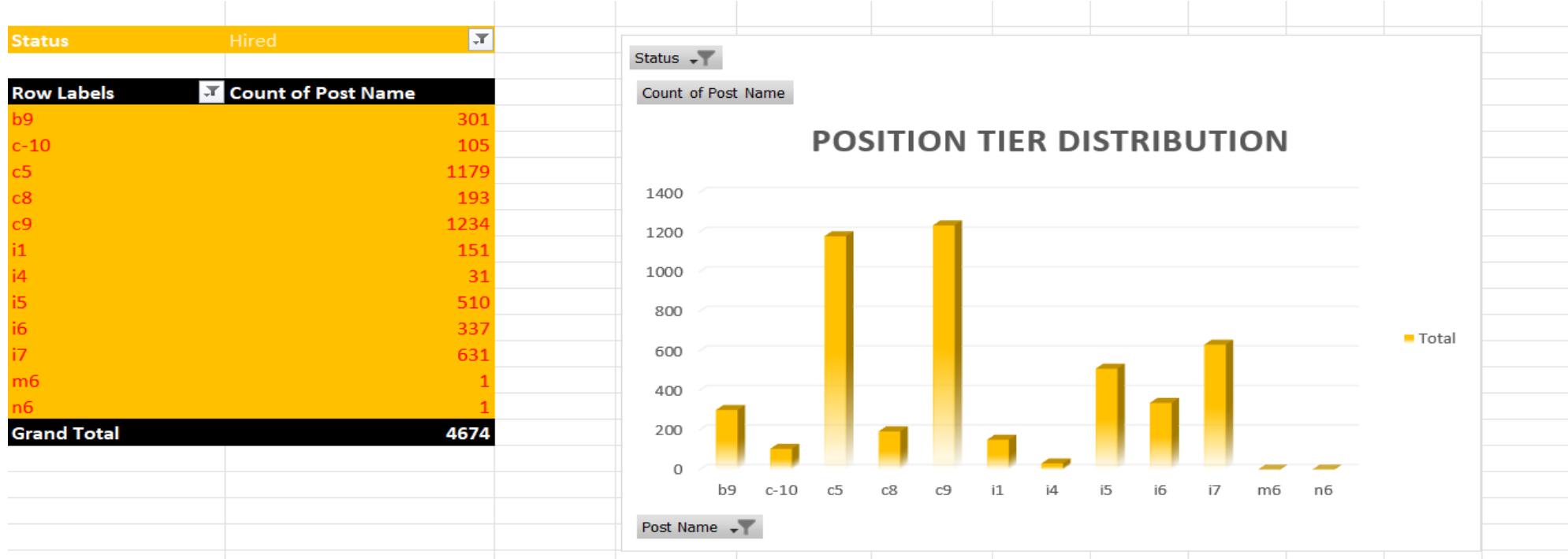
- Operations (39%)
- Service (29%)
- Other departments like Sales (10%), Production (5%) follow behind.

Operations and Service dominate the hiring, likely indicating company focus areas or workforce needs..



# Position Tier Distribution

- Most common roles: c5 (1179), c9 (1234), i7 (631)
- Some tiers (m6, n6) have minimal hires, suggesting rare or special positions.
- Clear hiring focus on a few select tiers. You could further explore job responsibilities per tier in the full report.

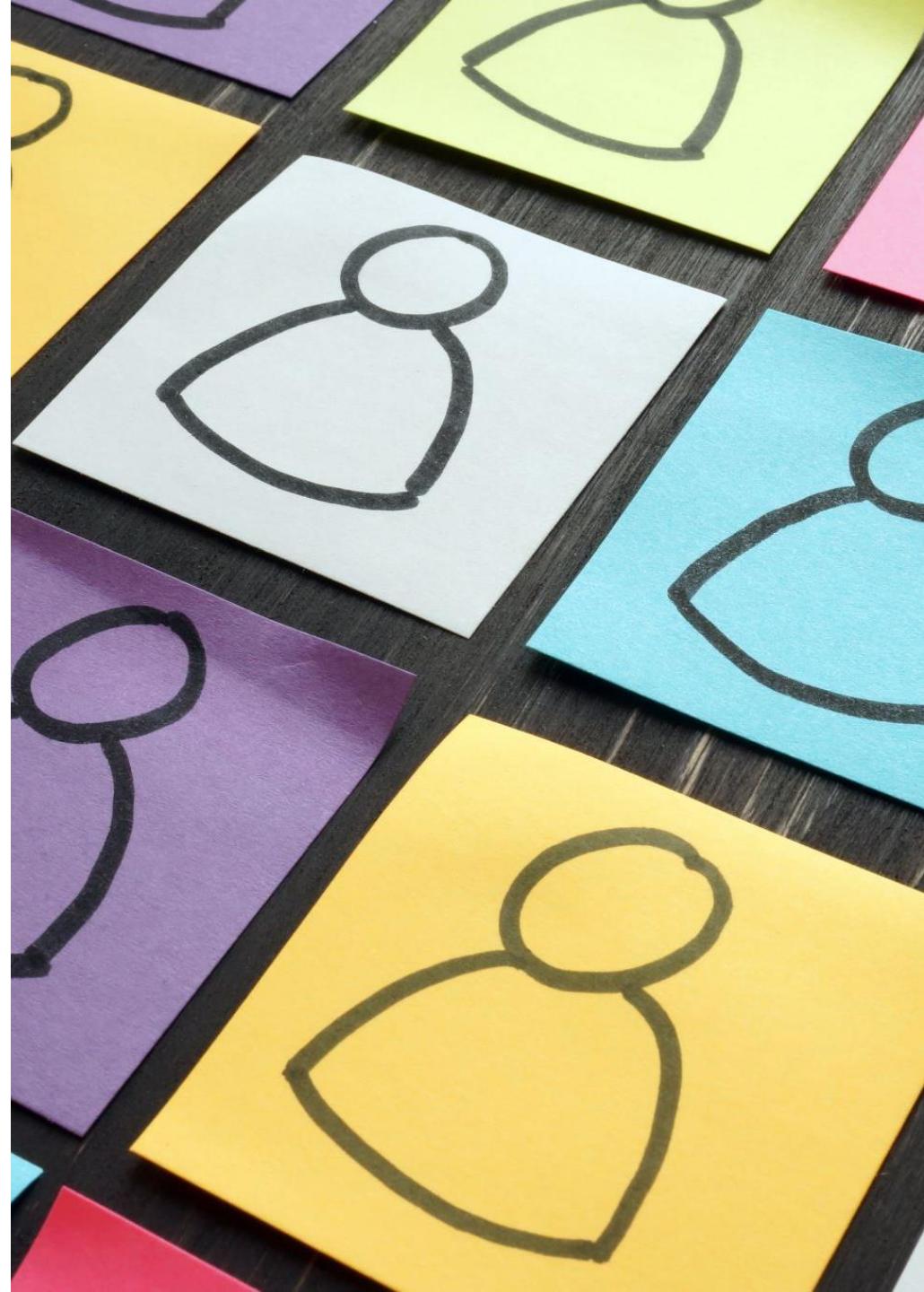


# Conclusion

- This analysis highlights key aspects of the hiring process.
- The company focuses majorly on operational roles with balanced salary offerings across genders.
- A few tiers dominate the hiring pattern, indicating preferred positions or recurring needs.
- These insights can guide recruitment strategy and workforce planning.

**Drive Link:**

[https://docs.google.com/spreadsheets/d/1KDySkzg2JgkjVab0P8\\_rZVifzMxmlB6/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1KDySkzg2JgkjVab0P8_rZVifzMxmlB6/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true)



## Learnings & Final Thoughts

This portfolio is more than a collection of projects - it's a reflection of my journey from curiosity to confidence in the world of data.

Over the course of this internship, I explored data like a storyteller, interrogated spreadsheets like a detective, and delivered insights like a strategist. Each project challenged me to think deeper, work smarter, and bridge the gap between numbers and narratives.

Through this journey, I learned that:

- **Clean data is powerful data.**
- **A well-crafted visualization can speak louder than a thousand rows.**
- **Asking the right question is just as important as finding the right answer.**

I've grown fluent in tools like **Excel, SQL, and Power BI**, but more importantly, I've learned how to use them to solve real business problems - whether it's reducing loan defaults, optimizing hiring processes, or identifying customer behavior trends.

As I close this chapter, I carry forward a sharpened skill set, a problem-solving mindset, and a deep respect for the impact of data. I'm excited to continue this journey - not just as a data analyst, but as a lifelong learner and storyteller of insights.