HR Analytics: Commute and Attrition

Natalie Pong
Feb 2025

As a recruiter, I often see companies struggle to attract candidates to onsite jobs, especially in Silicon Valley where remote work became more common since the pandemic. However, from a human resources standpoint, I'm curious if commute time really has any impact on the turnover rates. Do people who live further away from the office tend to have a higher turnover rate?

Although I don't have access to all the confidential HR data from every single company to generalize, I can investigate the employee retention factors of IBM using Kaggle's open source data. That would narrow down my question to **"Do people who live further away from the IBM office tend to have a higher turnover rate?"** I'm guessing that it does at a certain threshold, but let's test it out.

The database has key columns below:

Column	Type of Data
`EmployeeNumber`	Discrete numerical data that serves as a unique employee ID
`DistanceFromHome`	Continuous numerical data, which I assume is the number of minutes a person spends commuting one way from home to the IBM office
`Attrition`	Regular categorical data, "Yes" means that the person no longer works at IBM, and "No" means that the person currently works at IBM

To start, I opened up VS Code. Claude suggested SQLite because it supports CTE's and JOIN functions.

```
import pandas as pd
import sqlite3
import matplotlib.pyplot as plt
```

Then I uploaded Kaggle's IBM HR Excel file and called it **ibm_hr**

```
ibm_hr = pd.read_csv("/Users/nataliepong/Desktop/IBM_HR_Analytics.csv")
```

I use my computer for other memory-heavy files. So ChatGPT suggested RAM as the computationally frugal solution instead of creating a file to make a SQLite in-memory database.

```
conn = sqlite3.connect(":memory:")
```

Then I proceeded to load the dataframe into SQLite.

```
ibm_hr.to_sql("ibm_hr", conn, index=False, if_exists="replace")
```

HR Analytics: Commute and Attrition

Natalie Pong
Feb 2025

I used SQL bucketing to prepare data for a bar graph to show each employee's commute on the X-axis and number of employees on the Y-axis. I categorized the data by commute as well as employment status (current and ex-employees). Commute is categorized as the following:

- “Easy commute” is just 15 mins or less
- “Moderate commute” is over 15 mins but less than 30 mins
- “Far” is the category for commute over 30 mins

I also used CTE's for both current and ex-employees.

```
query = """
WITH
   current_employee AS
   (SELECT EmployeeNumber, DistanceFromHome
   FROM ibm_hr
   WHERE Attrition = 'No'
   ),
   ex_employee AS
   (SELECT EmployeeNumber, DistanceFromHome
   FROM ibm_hr
   WHERE Attrition = 'Yes'
   )

SELECT
CASE
   WHEN c.DistanceFromHome <= 15 THEN 'Easy commute'
   WHEN c.DistanceFromHome > 15 AND c.DistanceFromHome < 30 THEN 'Moderate commute'
   ELSE 'Far'
   END
   AS commute,
   'Current Employee' AS status,
 COUNT(*) AS employee_count
FROM current_employee AS c
GROUP BY commute

UNION ALL

SELECT
CASE
   WHEN e.DistanceFromHome <= 15 THEN 'Easy commute'
   WHEN e.DistanceFromHome > 15 AND e.DistanceFromHome < 30 THEN 'Moderate commute'
   ELSE 'Far'
   END
```

HR Analytics: Commute and Attrition

Natalie Pong
Feb 2025

```
   AS commute,
   'Ex-Employee' AS status,
 COUNT(*) AS employee_count
FROM ex_employee AS e
GROUP BY commute
"""
```

Then I ran SQL on the Pandas dataframe and called it **ibm_hr_df**

```
ibm_hr_df = pd.read_sql(query, conn)
conn.close()
```

In Python, I categorized the results by commute and IBM employment status.

```
for status in ibm_hr_df["status"].unique():
   subset = ibm_hr_df[ibm_hr_df["status"] == status]
   plt.bar(subset["commute"], subset["employee_count"], label=status)
```

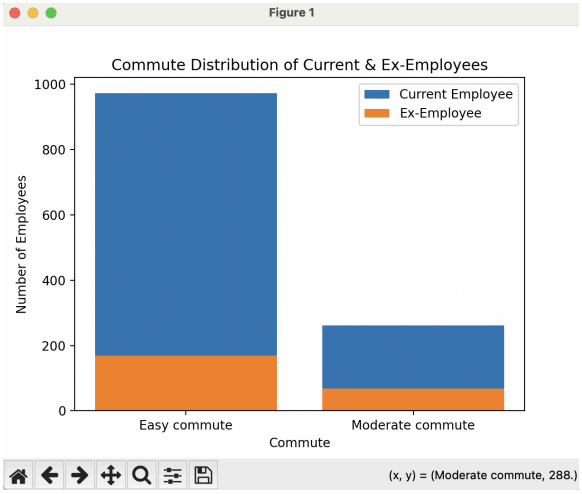
Then I used Python's matplotlib package to create the bar graph

```
plt.xlabel("Commute")
plt.ylabel("Number of Employees")
plt.title("Employee Commute Distribution of Current & Ex-Employees")
plt.legend()
plt.show()
```

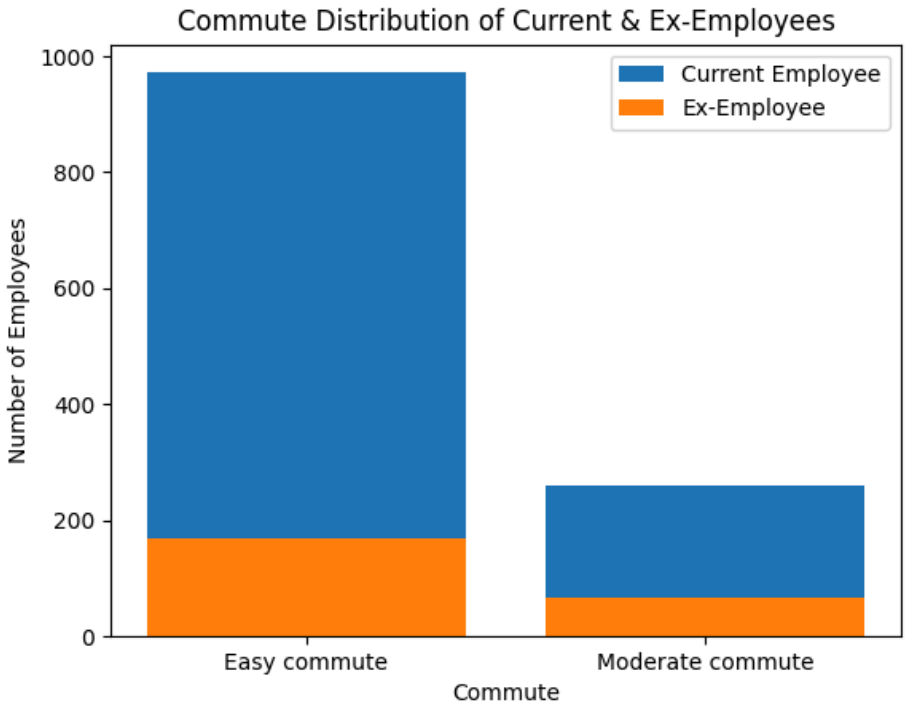
HR Analytics: Commute and Attrition

Natalie Pong
Feb 2025

VS Code gave the following graph as screenshotted:



The resulting graph that I can share with stakeholders would be:



Conclusion

Because I don't have context on how the data was collected, I can only make a causal statement instead of a generalization for all companies in the world. To answer the question, **"Do people who live further away from the IBM office tend to have a higher turnover rate?",** my hypothesis is incorrect. I can only tell stakeholders that in general, office commute has no relationship with attrition for employees in this dataset. However, it's important to note a

confounding variable like the type of commute, which is missing from this dataset. For example, commuting 30 mins by foot is considered “Far” compared with commuting 30 mins by car or train. Additionally, there’s a possibility that commute changes over time. For example, if a new hire signs an offer with their home address at the time of the offer, relocates for the job, and then buys a home in another neighborhood after getting promoted at IBM, their commute has changed over time. Moreover, this data might ignore a reverse causal effect, i.e. most employees were hired because they were close to the office. Afterall, they were well aware of the commute before signing the offer.

Further Consideration

Suppose the Human Resources Department (HR) requested this study to understand factors that impact attrition rates. Before jumping to conclusions with this study, they can perform exit interviews to dig deeper into the reasons for people leaving a company. Better work culture elsewhere, better work-life balance elsewhere, promotion elsewhere, getting fired vs getting laid off, relocation for a spouse’s job, career change, and higher compensation at another company are other possible reasons for leaving a company. I would also like to highlight that exit interview responses may not be entirely honest; employees who want to maintain a positive relationship with the company may sugarcoat their answers knowing that it stays in the company records after they leave.

If hiring managers requested this study to understand factors that impact hiring, they can use talent mapping to observe the locations of talents of interest, then decide on whether it makes sense to hire talent closer to the office. The “remote work” option can be considered if the candidate market is too far from the office, and if candidates are turning down interviews or job offers due to commute. However, we need to keep in mind that the job function may impact the choice of remote work. For example, a manufacturing engineer at IBM won’t be able to bring hardware to their home and efficiently and remotely collaborate with other manufacturing engineers from other locations. Imagine lugging hefty machines and assembly lines between homes! On the other hand, a recruiter at IBM can do their work remotely since they can speak to stakeholders over Zoom and manage candidate interviews and CV’s in an applicant tracking system (ATS).

For a more thorough study on the impact of commute on attrition, I would need job function data to prevent remote-friendly roles from skewing the data and aggregated sums of responses from exit interviews to rule out other potential variables.