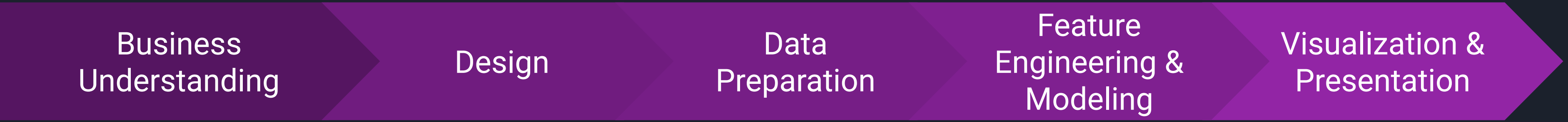
PMI Technical Assessment People Analytics Data Science Lead

Applicant

Aldo Cantu

Methodology



Business Understanding

- Understand and frame the problem
- Gather Requirements

Design

- Data collection
- Initial EDA & Data assessment
- Bucketing Strategies
- Making assumptions

Data Preparation

- Data Wrangling
- EDA

Feature Engineering & Modeling

- Statistical methods
- Feature Engineering
- Model building
- Model assessment

Visualization & Presentation

- Graphs / Plot selection
- Make recommendations

Technical challenge:
San Francisco Taxi Cabs 🚕

San Francisco Taxi Cabs

Problem Statement:

Taxis that operate in San Francisco often roam without passengers. This yields unnecessary CO_2 emissions that affect our current global warming situation. The goal of this challenge is to analyse and calculate these emissions based on information collected of about 500 taxi cabs on a 30 day time frame.

Specified Assumptions:

⛽🔜🔋 Monthly taxi fleet change rate = 15%
☁️ Average CO_2 emission of a passenger vehicle = 400 gr./m

Requirements:

- Calculate potential yearly reduction in CO_2 emissions caused by taxi cabs without passengers.
- Data Visualization

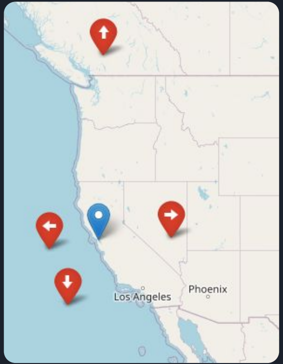
Bonus:

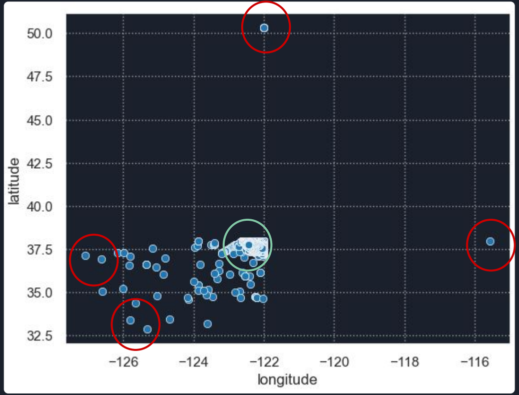
- 📍 Predict the next place a passenger will hail a cab

Initial EDA, Data assessment & Assumptions

Location Analysis

As-is





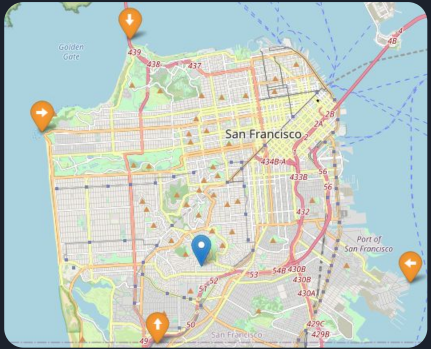
💡 Assumptions

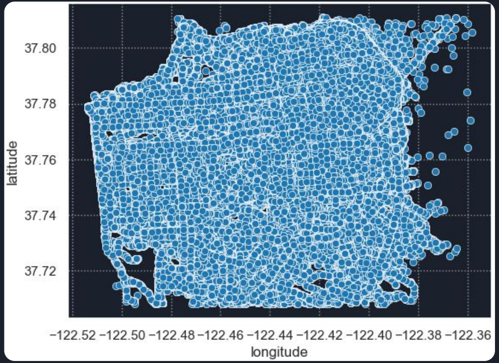
Farmost Cardinal Point	Limit
West Longitude	-122.51494
East Longitude	-122.357034
North Latitude	37.811054
South Latitude	37.7080955

Hard Limits (Physical)!

OUTLIERS
106275 Fares outside the assumed limits

Correction

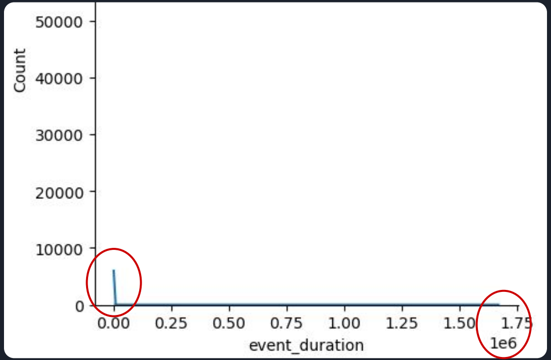




Initial EDA, Data assessment & Assumptions

Event Duration Analysis

As-is

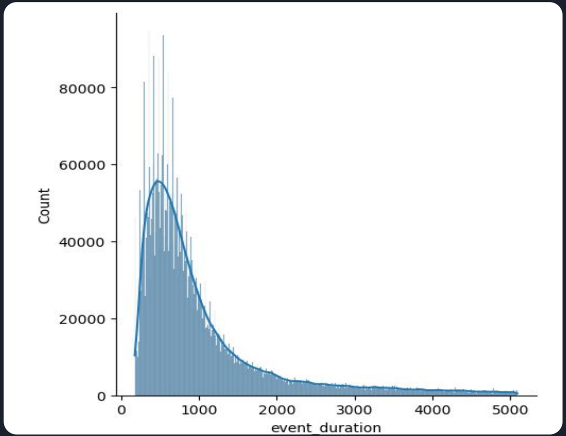


💡 Assumptions

occupancy	min duration	max duration	Quantile
empty	179 sec	5090 sec.	5/75
occupied	242 sec	2039 sec.	5/95

Using Quantiles!

Correction



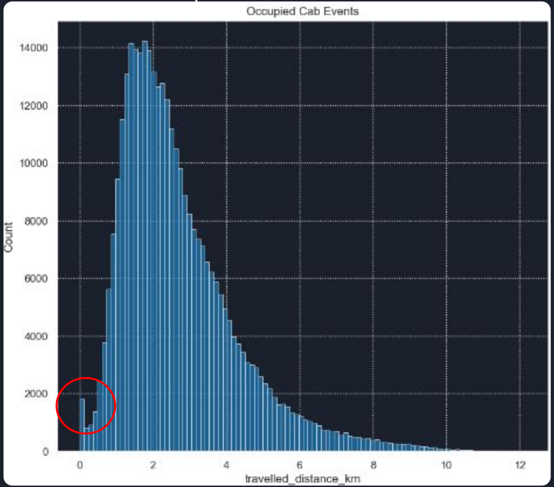
OUTLIERS
235206 Fares outside the assumed limits

Initial EDA, Data assessment & Assumptions

Distance Analysis

As-is

Occupied Cab Events



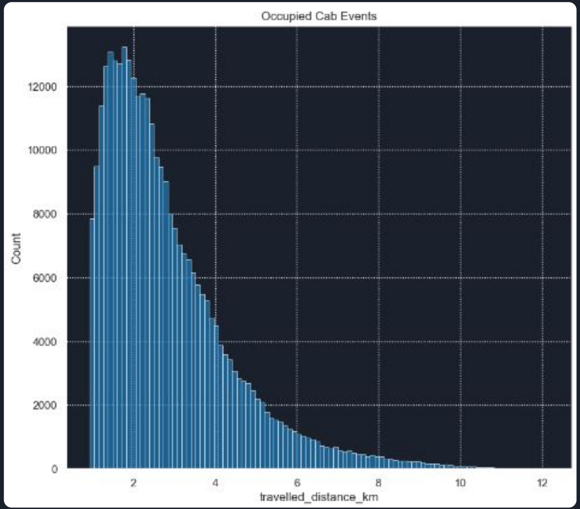
💡 Assumptions

Occupied cabs that travel less than 0.92 km are outliers (non-inclusive)

5% Quantile

Using Quantiles!

Correction



OUTLIERS
21046 Fares outside the assumed limits

Potential yearly reduction in CO_2 emissions caused by taxi cabs without passengers

empty_cabs_ds

Travelled Distance
394,141 km

Time Period
23.97 Days

Accounting for Leap year

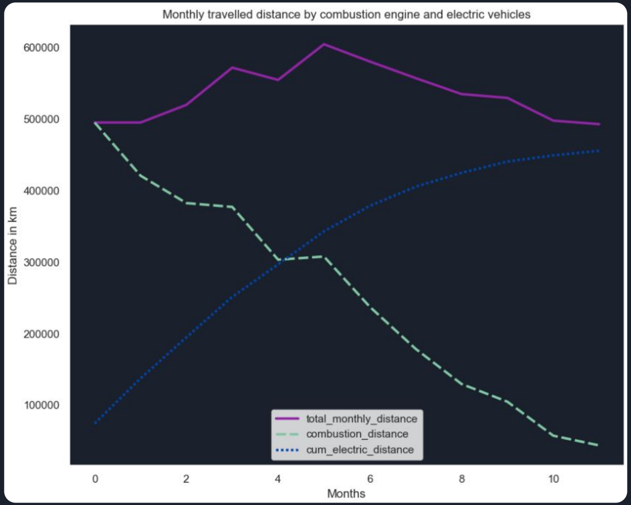
Time Scale
30,4375 Days

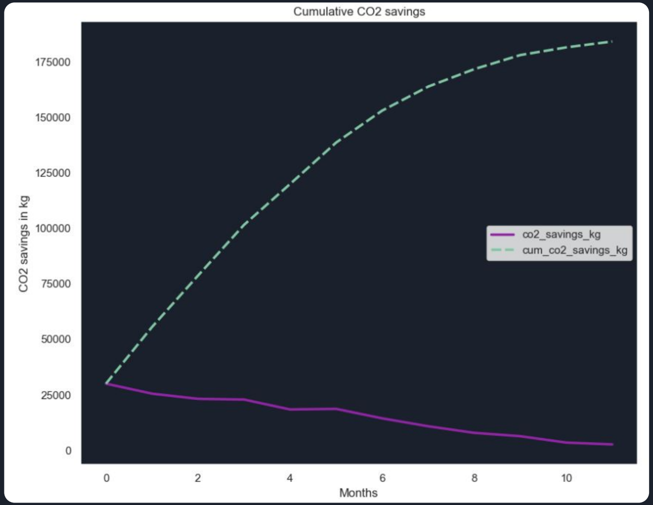
Travelled Distance
500,361 km
310, 909.9 mi

12 Month loop
💡 10% travelled distance variability

⚠️ Given chosen assumptions

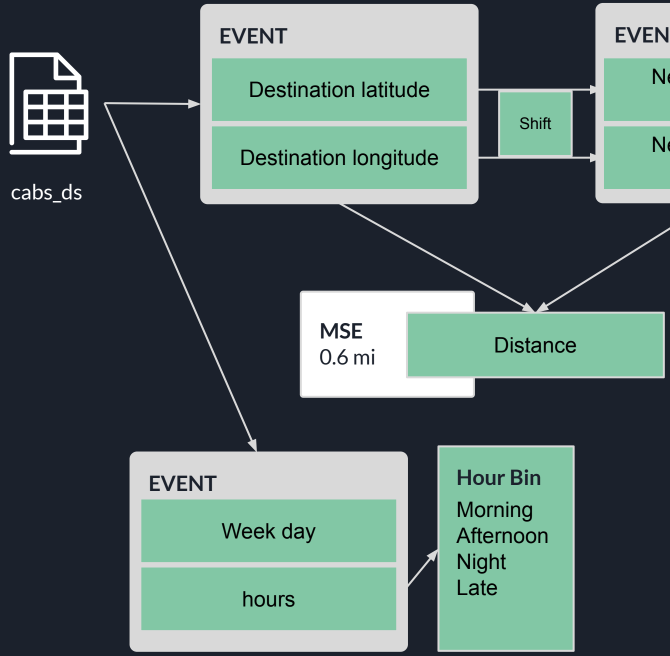
ANSWER
CO_2 ~1557 Ton





San Francisco Taxi Cabs

Next Fare Predictor



cabs_ds

EVENT

Destination latitude

Destination longitude

Shift

EVENT

Next event origin latitude

Next event origin latitude

MSE
0.6 mi

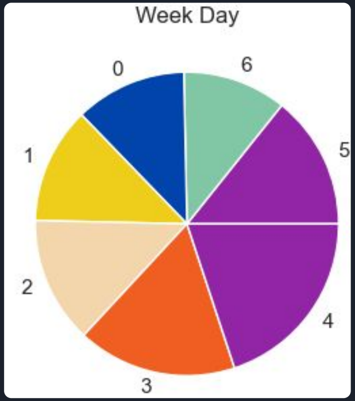
Distance

EVENT

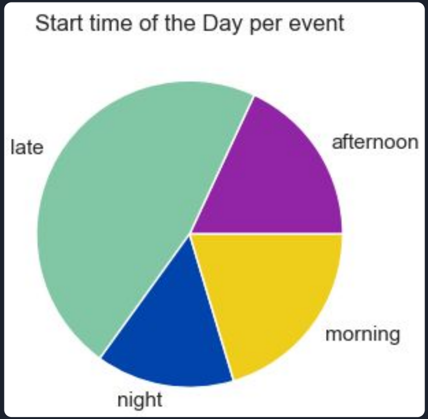
Week day

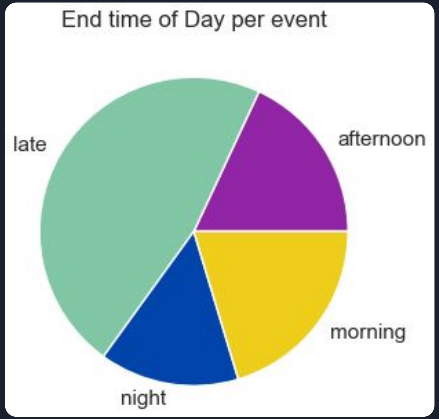
hours

Hour Bin
Morning
Afternoon
Night
Late



KMeans



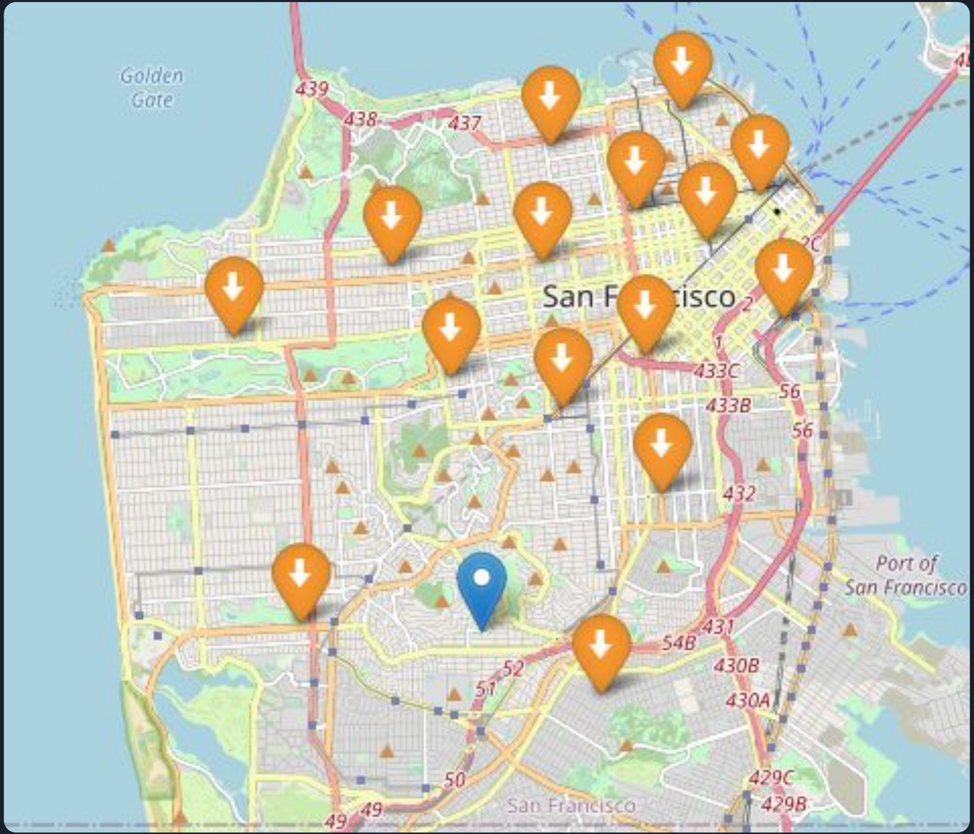


🟩 Features

EVENT = state change (free <-> occupied)

🚕San Francisco Taxi Cabs

Next Fare Predictor



MSE
1.64 mi

Distance

Further Steps

- Adding a seasonality
- Taking weather into account
- Further time decomposition
- Implement a cross entropy Naive Bayes or KNN model with multivariate grid-system

Use case:
Turnover Analysis 👎

Turnover Analysis

Problem Statement:

The CHRO has asked the Head of Talent Management to devise a retention strategy to make sure top talent is retained. The goal is to analyze employee turnover data that has been collected over the past twelve months. To answer the following questions:

1. Is the company losing top talent?
2. If so, why and how can current top talent be retained

Requirements:

- Data Visualization
- Generate a strategy that will help lessen the turnover of top talented employees

Reframing the problem

Since the question “*Is the company losing top talent?*” is not a testable hypothesis a reframing of the problem needs to be done.

On a study done by an employee retention platform [1], the overall average turnover rate in the US is about **47.2%**.

Moreover, another article written by Michael Mankins (2017)[2] on the Harvard Business Review site:
*“On average, **15%** of a company’s workforce - roughly one in seven employees - are A player or “stars”.*

So our hypothesis stands by the following:

“Is the company losing more or equal to 6.7% of top talent?”

👎 Turnover Analysis

Is the company losing top talent?

$$TurnoverRate(\%) = \frac{\#Separations}{Avg.\#Employees} * 100$$

* High potentials numbers used for High Potential turnover

	Company		Global
Overall company’s turnover rate:	11.2%	👍	47.2%
High Potentials turnover rate:	16.25%	👎	6.7%

Yes, the company is losing top talent with a significant number compared to what other companies have previously reported.

Data assessment

From the initial data state, a few remarks have surfaced:

⚠️ "Tenure" and "Time in latest role" **features have negative values**

⚠️ Max values for "Tenure" and "Time in latest role" are **above 90 years** whereas realistically a maximum value of 50 years should be addressed

⚠️ "Actual monthly pay" feature seems to fit the description and values of a **week's worth pay**

⚠️ In "Country" and "Nationality" feature, *Monrovia* is declared as country/nationality whereas the correct value is *Liberia*

⚠️ In "Race" feature, "White" and "white" (upper- & lowercase) is treated as different category

⚠️ There is a typo on the "Sexuality" feature. Instead of *Non-hetersexual* it should be *Non-heterosexual*. (missing an "o")

⚠️ There is a typo on the "Career opportunities Survey score" feature. It should be "Career Opportunities Survey Score"

Assumptions

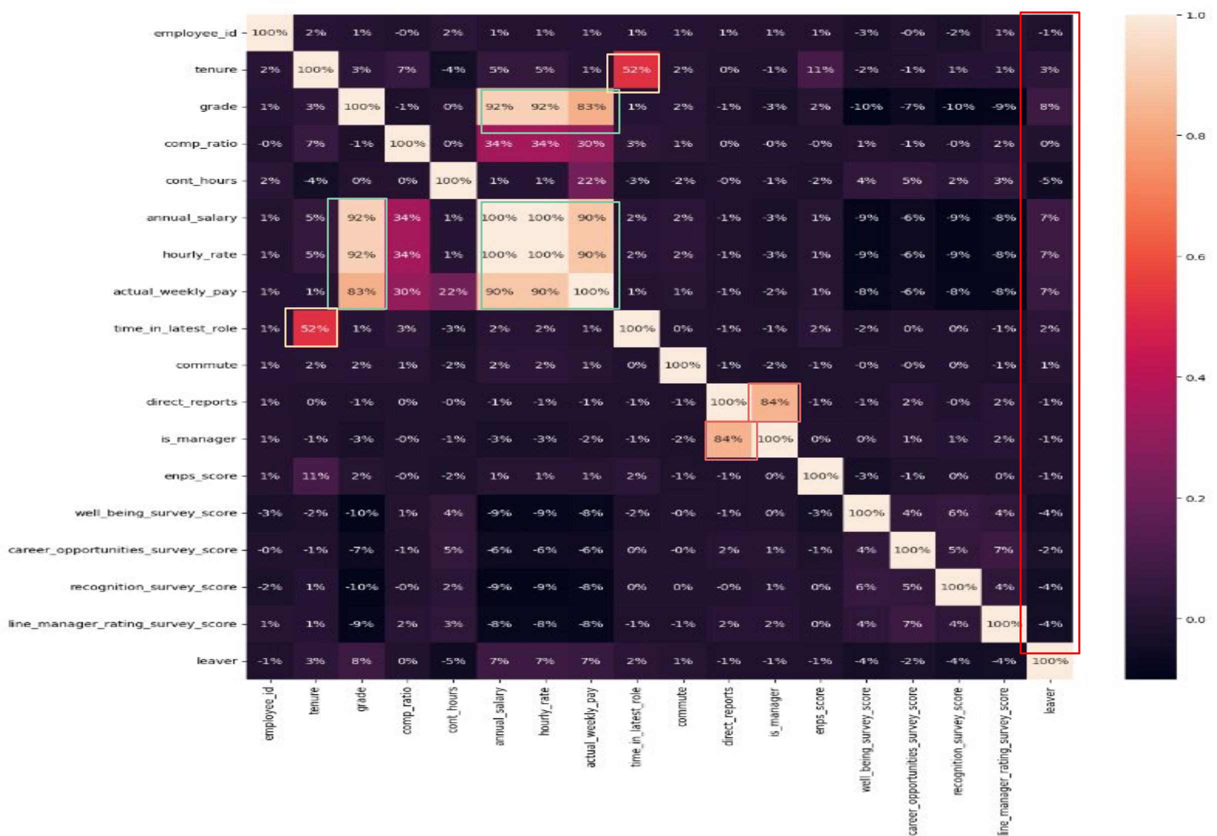
- Negative values are filtered out
- Maximum tenure of 50 years

👎 Turnover Analysis

Initial EDA

Pearson Correlation

Correlation matrix of Employee Turnover Data



	employee_id	tenure	grade	comp_ratio	cont_hours	annual_salary	hourly_rate	actual_weekly_pay	time_in_latest_role	commute	direct_reports	is_manager	enps_score	well_being_survey_score	career_opportunities_survey_score	recognition_survey_score	line_manager_rating_survey_score	leaver
employee_id	100%	2%	1%	-0%	2%	1%	1%	1%	1%	1%	1%	1%	1%	-3%	-0%	-2%	1%	-1%
tenure	2%	100%	3%	7%	-4%	5%	5%	1%	52%	2%	0%	-1%	11%	-2%	-1%	1%	1%	3%
grade	1%	3%	100%	-1%	0%	92%	92%	83%	1%	2%	-1%	-3%	2%	-10%	-7%	-10%	-9%	8%
comp_ratio	-0%	7%	-1%	100%	0%	34%	34%	30%	3%	1%	0%	-0%	-0%	1%	-1%	-0%	2%	0%
cont_hours	2%	-4%	0%	0%	100%	1%	1%	22%	-3%	-2%	-0%	-1%	-2%	4%	5%	2%	3%	-5%
annual_salary	1%	5%	92%	34%	1%	100%	100%	90%	2%	2%	-1%	-3%	1%	-9%	-6%	-9%	-8%	7%
hourly_rate	1%	5%	92%	34%	1%	100%	100%	90%	2%	2%	-1%	-3%	1%	-9%	-6%	-9%	-8%	7%
actual_weekly_pay	1%	1%	83%	30%	22%	90%	90%	100%	1%	1%	-1%	-2%	1%	-8%	-6%	-8%	-8%	7%
time_in_latest_role	1%	52%	1%	3%	-3%	2%	2%	1%	100%	0%	-1%	-1%	2%	-2%	0%	0%	-1%	2%
commute	1%	2%	2%	1%	-2%	2%	2%	1%	0%	100%	-1%	-2%	-1%	-0%	-0%	0%	-1%	1%
direct_reports	1%	0%	-1%	0%	-0%	-1%	-1%	-1%	-1%	-1%	100%	84%	-1%	-1%	2%	-0%	2%	-1%
is_manager	1%	-1%	-3%	-0%	-1%	-3%	-3%	-2%	-1%	-2%	84%	100%	0%	0%	1%	1%	2%	-1%
enps_score	1%	11%	2%	-0%	-2%	1%	1%	1%	2%	-1%	-1%	0%	100%	-3%	-1%	0%	0%	-1%
well_being_survey_score	-3%	-2%	-10%	1%	4%	-9%	-9%	-8%	-2%	-0%	-1%	0%	-3%	100%	4%	6%	4%	-4%
career_opportunities_survey_score	-0%	-1%	-7%	-1%	5%	-6%	-6%	-6%	0%	-0%	2%	1%	-1%	4%	100%	5%	7%	-2%
recognition_survey_score	-2%	1%	-10%	-0%	2%	-9%	-9%	-8%	0%	0%	-0%	1%	0%	6%	5%	100%	4%	-4%
line_manager_rating_survey_score	1%	1%	-9%	2%	3%	-8%	-8%	-8%	-1%	-1%	2%	2%	0%	4%	7%	4%	100%	-4%
leaver	-1%	3%	8%	0%	-5%	7%	7%	7%	2%	1%	-1%	-1%	-1%	-4%	-2%	-4%	-4%	100%

Initial EDA

A quick study of the
correlation matrix leads us to the following insights:

From the initial assessment of the information there are no significant correlations.

Some relationships are appreciated:
- Grade -> Annual Salary, Hourly Rate, Actual weekly pay
- Is Manager -> Direct Reports
- Tenure -> Time in latest Role

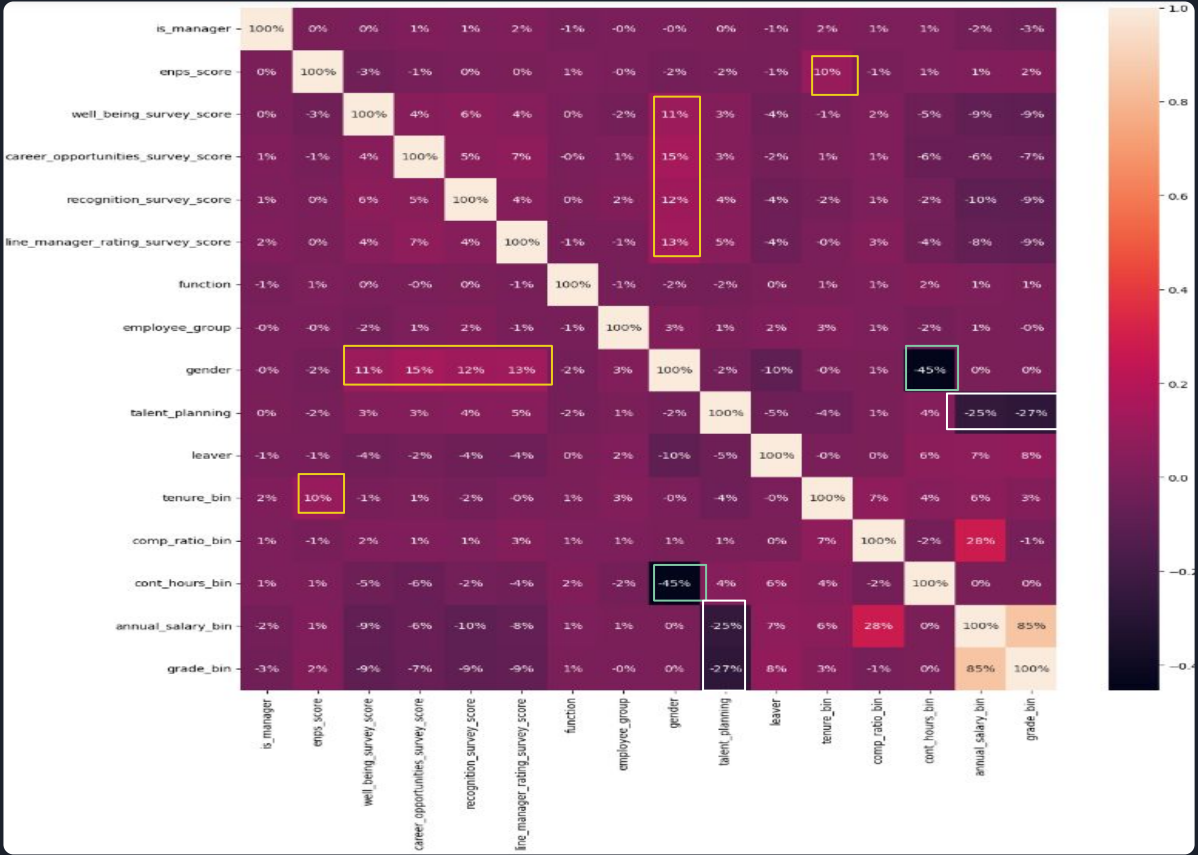
Further investigation should be done.

Business Understanding → Design → Data Preparation → Feature Engineering & Modeling → Visualization & Presentation

👎 Turnover Analysis

Exploratory Data Analysis

Correlation matrix of cleaned Employee Turnover Data



	is_manager	enps_score	well_being_survey_score	career_opportunities_survey_score	recognition_survey_score	line_manager_rating_survey_score	function	employee_group	gender	talent_planning	leaver	tenure_bin	comp_ratio_bin	cont_hours_bin	annual_salary_bin	grade_bin
is_manager	100%	0%	0%	1%	1%	2%	-1%	-0%	-0%	0%	-1%	2%	1%	1%	-2%	-3%
enps_score	0%	100%	-3%	-1%	0%	0%	1%	-0%	-2%	-2%	-1%	10%	-1%	1%	1%	2%
well_being_survey_score	0%	-3%	100%	4%	6%	4%	0%	-2%	11%	3%	-4%	-1%	2%	-5%	-9%	-9%
career_opportunities_survey_score	1%	-1%	4%	100%	5%	7%	-0%	1%	15%	3%	-2%	1%	1%	-6%	-6%	-7%
recognition_survey_score	1%	0%	6%	5%	100%	4%	0%	2%	12%	4%	-4%	-2%	1%	-2%	-10%	-9%
line_manager_rating_survey_score	2%	0%	4%	7%	4%	100%	-1%	-1%	13%	5%	-4%	-0%	3%	-4%	-8%	-9%
function	-1%	1%	0%	-0%	0%	-1%	100%	-1%	-2%	-2%	0%	1%	1%	2%	1%	1%
employee_group	-0%	-0%	-2%	1%	2%	-1%	-1%	100%	3%	1%	2%	3%	1%	-2%	1%	-0%
gender	-0%	-2%	11%	15%	12%	13%	-2%	3%	100%	-2%	-10%	-0%	1%	-45%	0%	0%
talent_planning	0%	-2%	3%	3%	4%	5%	-2%	1%	-2%	100%	-5%	-4%	1%	4%	-25%	-27%
leaver	-1%	-1%	-4%	-2%	-4%	-4%	0%	2%	-10%	-5%	100%	-0%	0%	6%	7%	8%
tenure_bin	2%	10%	-1%	1%	-2%	-0%	1%	3%	-0%	-4%	-0%	100%	7%	4%	6%	3%
comp_ratio_bin	1%	-1%	2%	1%	1%	3%	1%	1%	1%	1%	0%	7%	100%	-2%	28%	-1%
cont_hours_bin	1%	1%	-5%	-6%	-2%	-4%	2%	-2%	-45%	4%	6%	4%	-2%	100%	0%	0%
annual_salary_bin	-2%	1%	-9%	-6%	-10%	-8%	1%	1%	0%	-25%	7%	6%	28%	0%	100%	85%
grade_bin	-3%	2%	-9%	-7%	-9%	-9%	1%	-0%	0%	-27%	8%	3%	-1%	0%	85%	100%

Exploratory Data Analysis

After some refactoring of the data the following information could be concluded:

Weak Negative Relationships:
-Talent Planning -> Annual Salary, Grade

Moderate Negative Relationship:
-Cont. Hours -> Gender

Weak positive Relationships:
-Survey Scores -> Gender
-eNPS Score -> Tenure

Not in scope:
-Comp. Ratio -> Annual Salary

Now a breakdown of personnel marked as “High Potential” from the Talent Planning Dimension

Exploratory Data Analysis

Bucketing Strategies

💡 Since “Financial features” like Annual Salary, Hourly Rate and Actual week pay are fairly similar, only Annual Salary will be held on focus and further **bucketed in 10 thousand unit increments.**

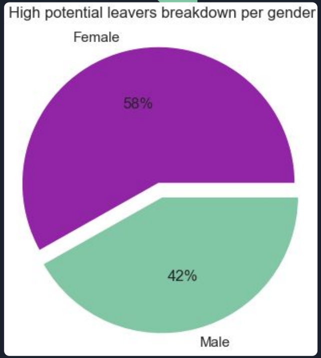
💡 Grade is also going to be grouped in **increments of 5**

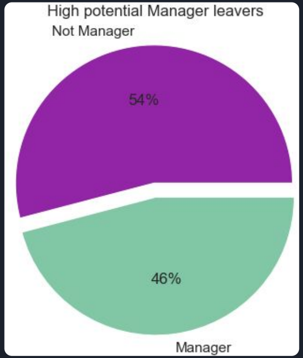
❌Ignored Features:

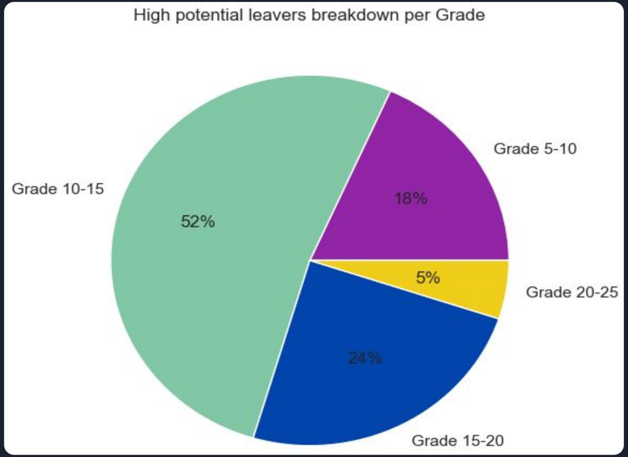
- Employee ID
- Hourly Rate
- Actual Week Pay
- Time in latest Role
- Commute
- Direct Reports
- Country
- Nationality
- Race
- Sexuality

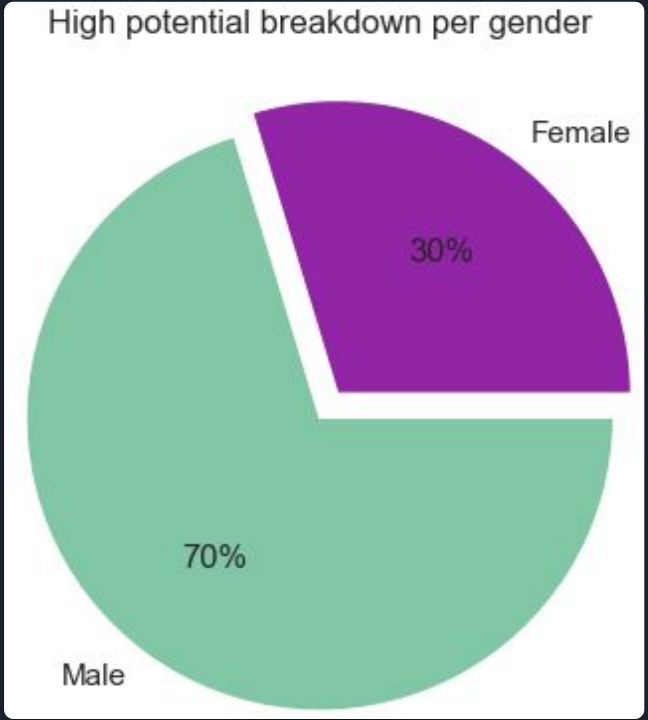
👎 Turnover Analysis

Visualization


High potential leavers breakdown per gender


High potential Manager leavers


High potential leavers breakdown per Grade


High potential breakdown per gender

⚠️A significant amount of people in management position are more likely to leave.

⚠️High potential females are more likely to leave than males

⚠️A significant amount of people in management position are more likely to leave.

⚠️Individuals within grade 10 to 15 tend to leave more frequently than the rest

⚠️Females are less prone to be classified as High Potential prospects.

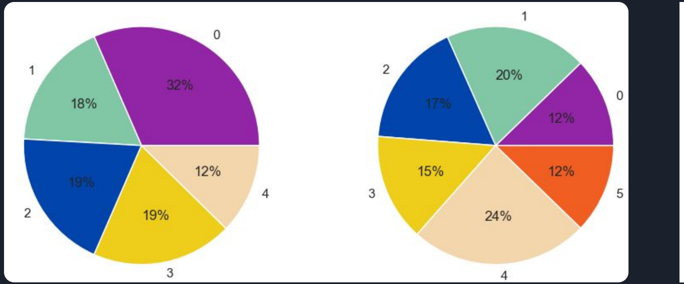
👎 Turnover Analysis

Visualization

Gender Comparison

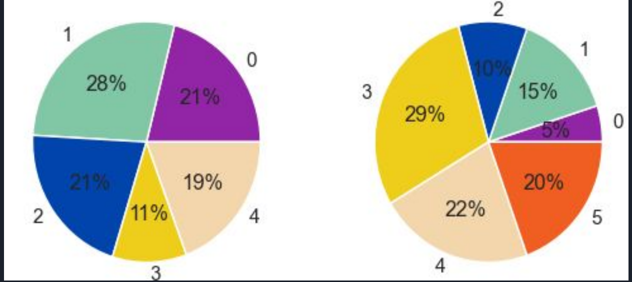
High potential leavers breakdown per **Well-being**

Female Male



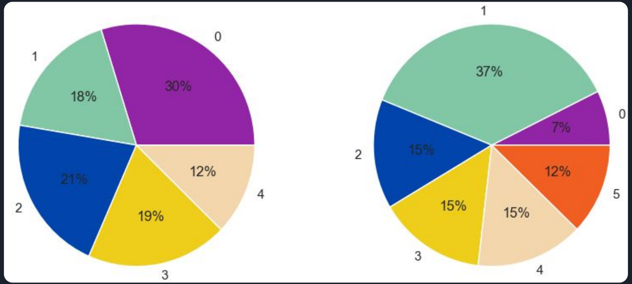
High potential leavers breakdown per **Career Opportunities**

Female Male



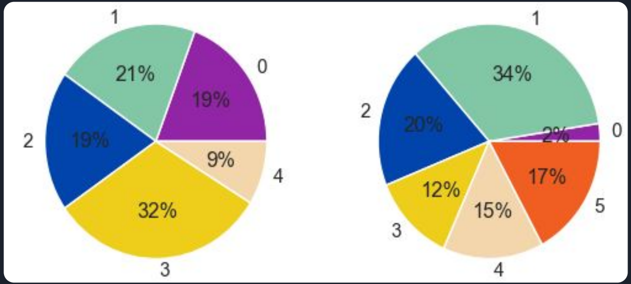
High potential leavers breakdown per **Line Manager Rating**

Female Male



High potential leavers breakdown per **Recognition**

Female Male



⚠️Females reported three times a value of Zero more than males

⚠️Males are two times more likely to report a “Very Good” well being score

⚠️Female have lower career opportunities compared to Males

⚠️Overall Managers are not enabling personnel to be successful on their roles

⚠️Males reported three times better recognition for their performance

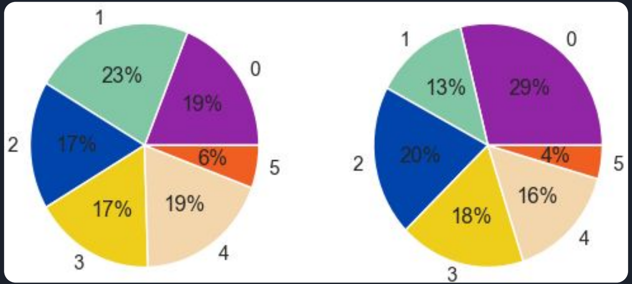
👎 Turnover Analysis

Visualization

Manager Comparison

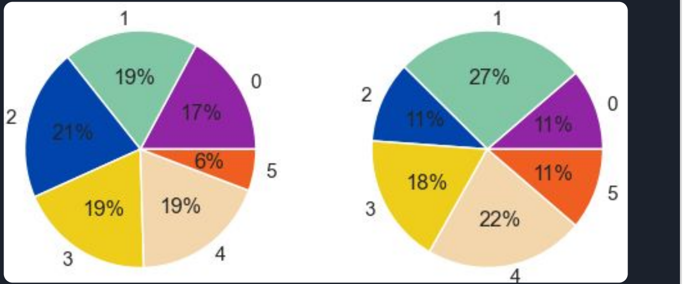
High potential leavers breakdown per **Well-being**

Not Manager Manager



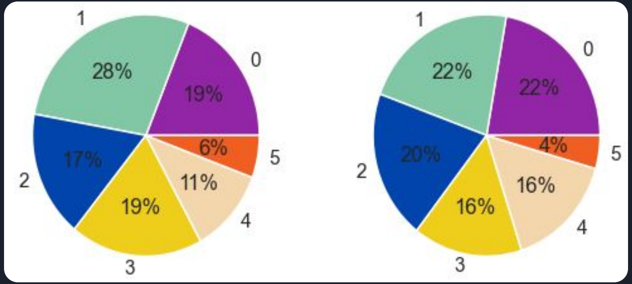
High potential leavers breakdown per **Career Opportunities**

Not Manager Manager



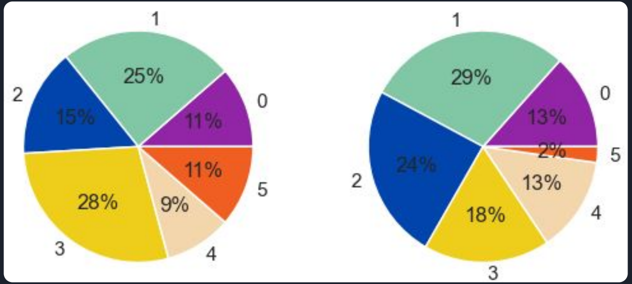
High potential leavers breakdown per **Line Manager Rating**

Not Manager Manager



High potential leavers breakdown per **Recognition**

Not Manager Manager



⚠️Majority of manager reported less than poor “Well-being” scores

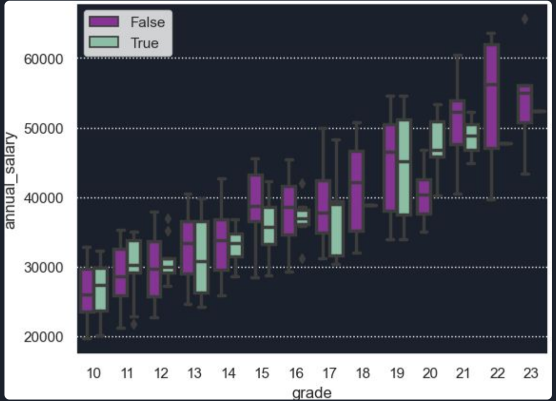
⚠️Majority reported less than poor “Line Manager Ratings”

⚠️Two thirds of overall High potential leavers reported less than poor “Career Opportunities” scores

⚠️Two thirds of overall High potential leavers reported poor “Recognition” scores

Visualization

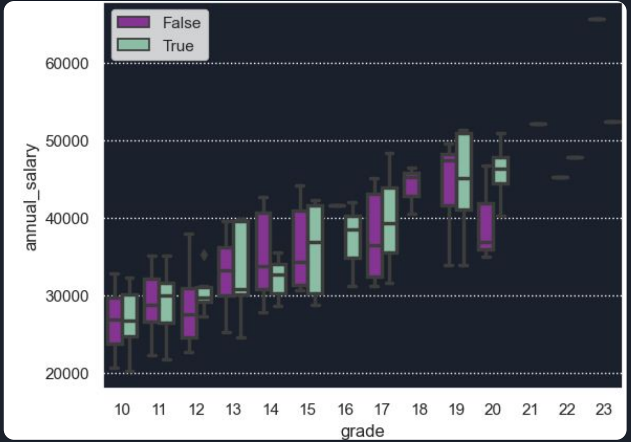
Comparison of grade and annual salary dimensioned by Leaver status of High Potential people



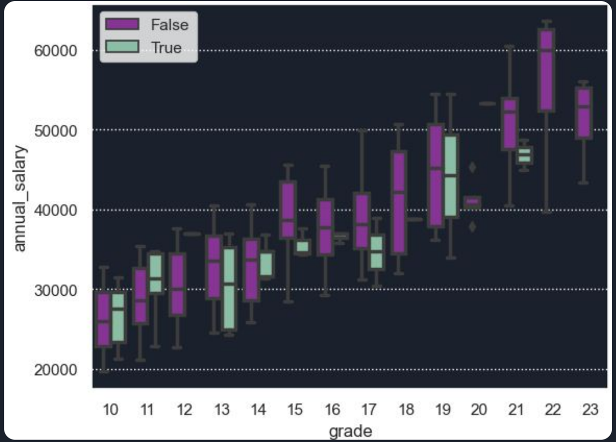
⚠️High potential people within grades 10, 13 and 19 are more likely to leave

⚠️Females within grade range from 10 to 19 tend to leave more often than males

Female

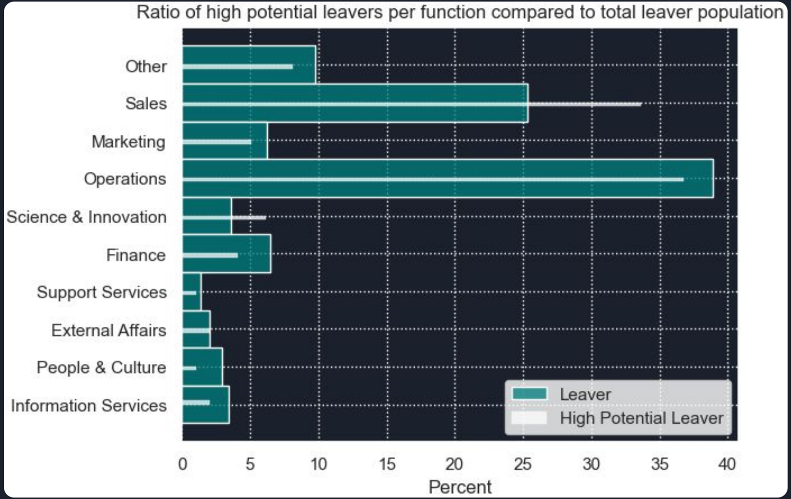


Male



👎 Turnover Analysis

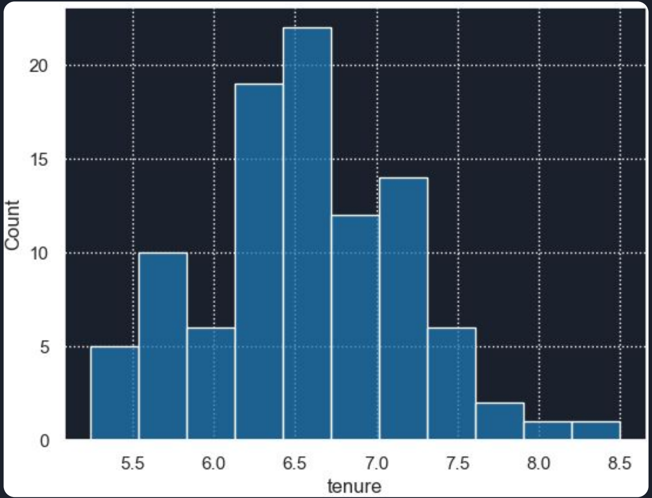
Visualization



⚠️Sales and Science & Innovation functions report higher turnover rates compared to the overall turnover numbers

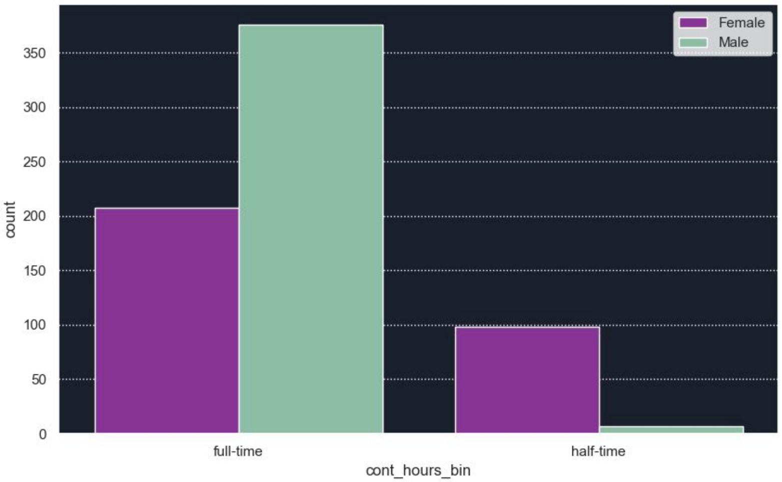
⚠️People between 5.5 and 7.5 Years of Tenure are more likely to leave the company

Attrition of High Potential individuals by Tenure



Visualization

Leaver count of Contracted Hours Type per Gender

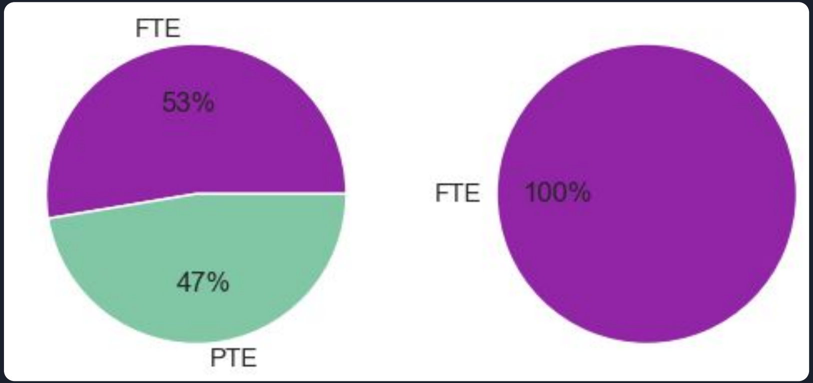


⚠️Males tend to predominate the full-time employment contracts
⚠️Females tend to leave the company a

High potential leavers by Contracted Hours breakdown per Gender

Female

Male



How can current top talent be retained

👍Since a significant part of people leaving the company (incl. managers) reside in the "mastered" years of tenure (between 5 and 10), a recommendation is to keep these resources **well tested** and look into **changing functions** to transfer their knowledge and skills exposing them to new challenges, teams and environments

👍Overall high potential people leaving the company reported poor career opportunities and recognition. The company could offer **changing career paths** and **planning events more often** to give high talent the chance to be recognized by their peers

👍High potential managers tend to leave more often due to poor "Well-being". One action is to offer a range of **extracurricular activities and work-life balance plans.**

👍Offering **family-oriented advantages** could help mitigate high potential female employees. **Better parental leave plans, position retention after birth and daycare** are but a few ideas.

Further Steps

Not all relationships were explored thoroughly, the following points would've been further analyzed:

- "Tenure" and "Time in latest role" correlations on turnover
- Overall data comparison (was heavily focused on high potential leavers)
- How hourly rate affects attrition instead of the used "Annual Salary"

Implemented a Random Forest Classifier (not shown in presentation)

48% of businesses say their **high-quality hires** come from **employee referrals**.[3]
(Linkedin)