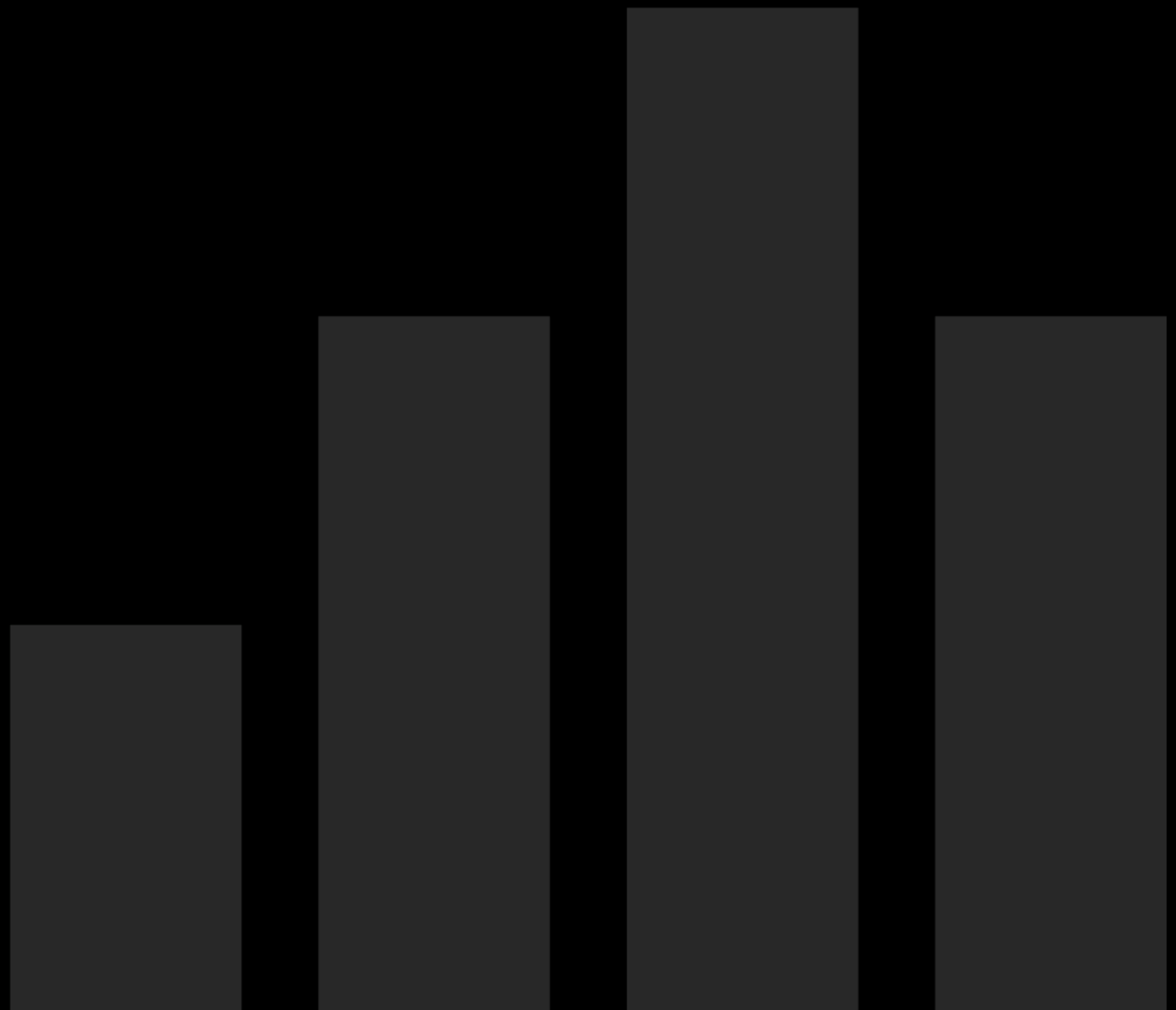
INTERNSHIP REPORT

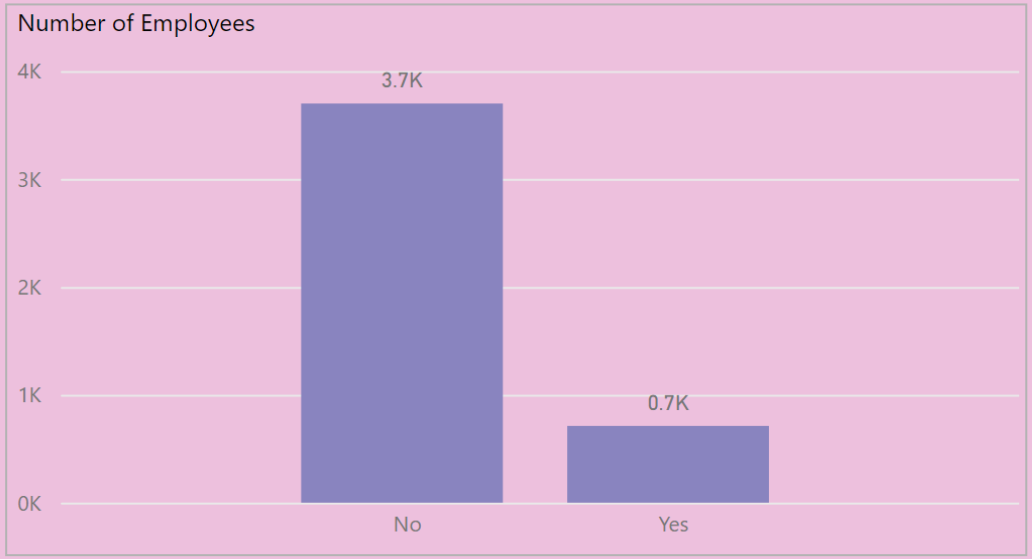
ELEVATE LABS

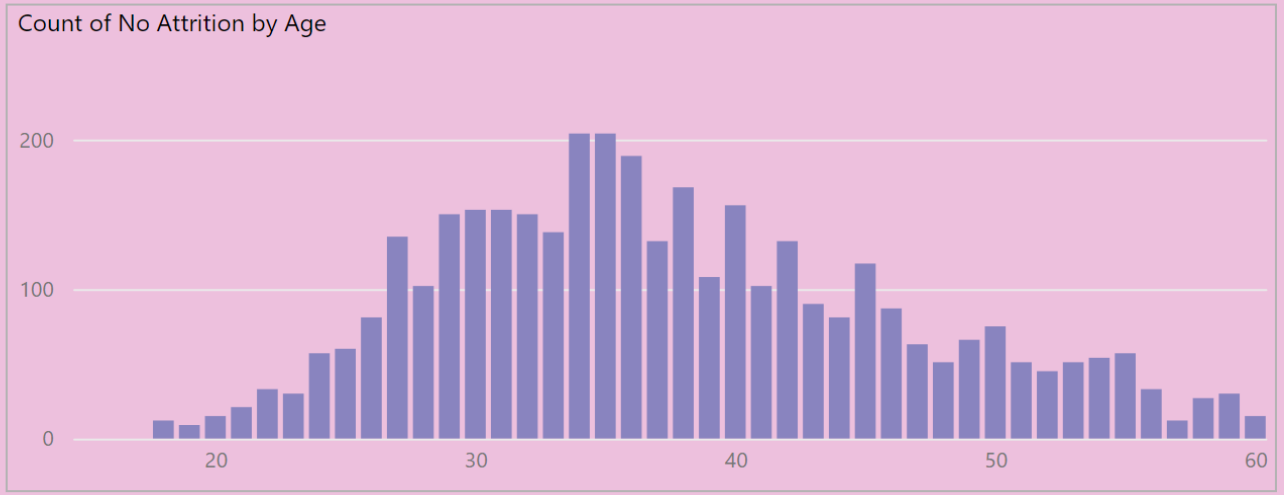
BY NAZIREEN SANIA

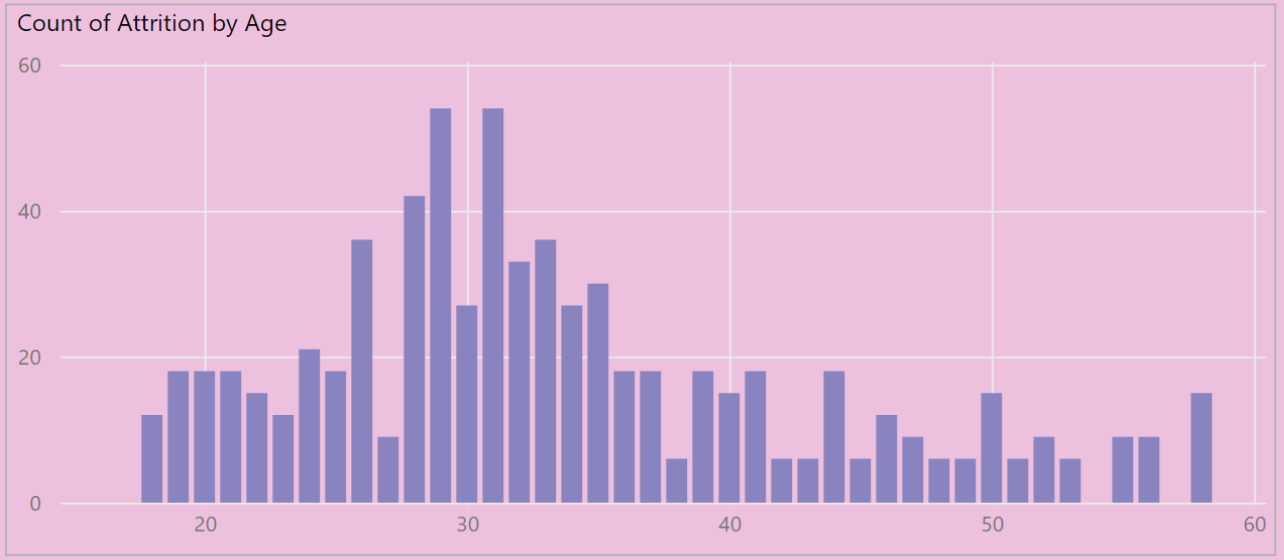
Employee Attrition

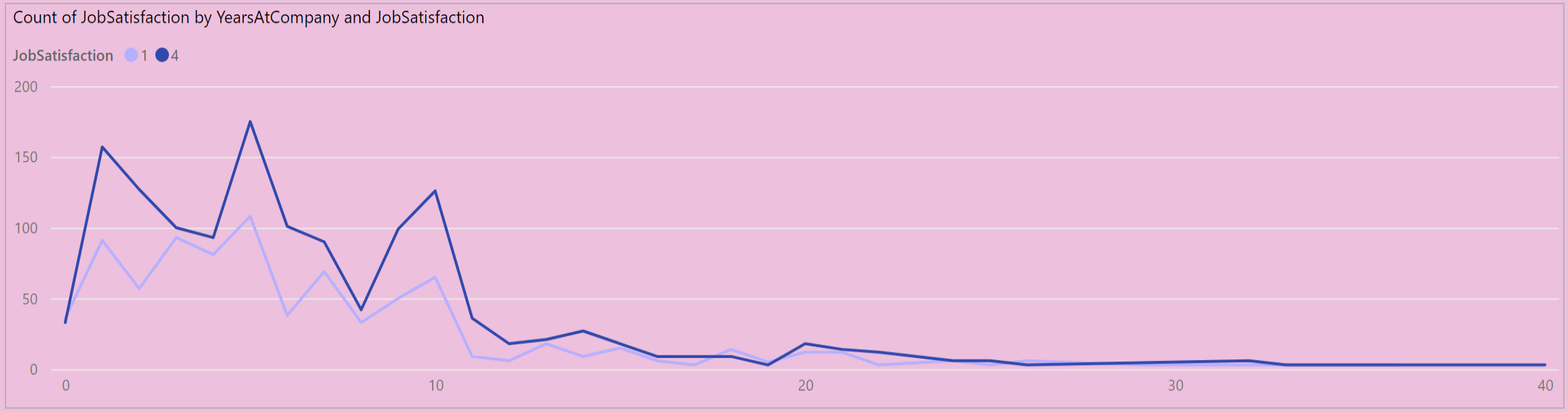
View in Power BI ↗

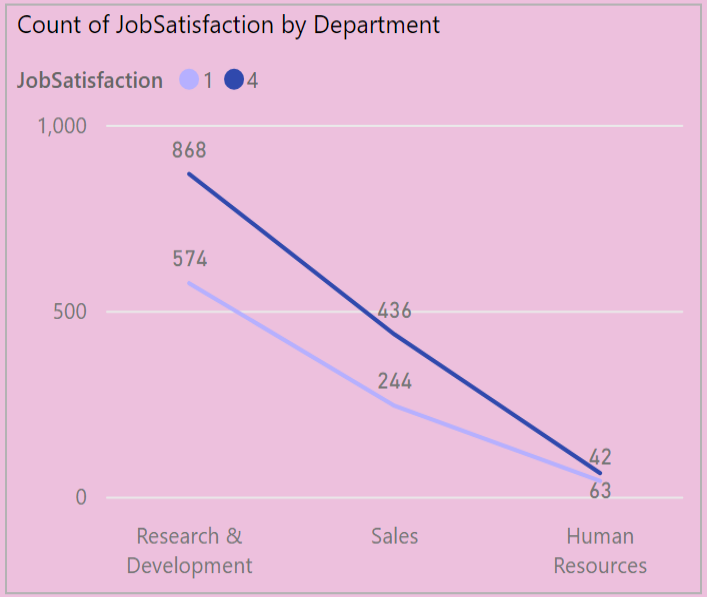



Number of Employees
4K
3K
2K
1K
0K
3.7K
0.7K
No
Yes


Count of No Attrition by Age
200
100
0
20
30
40
50
60

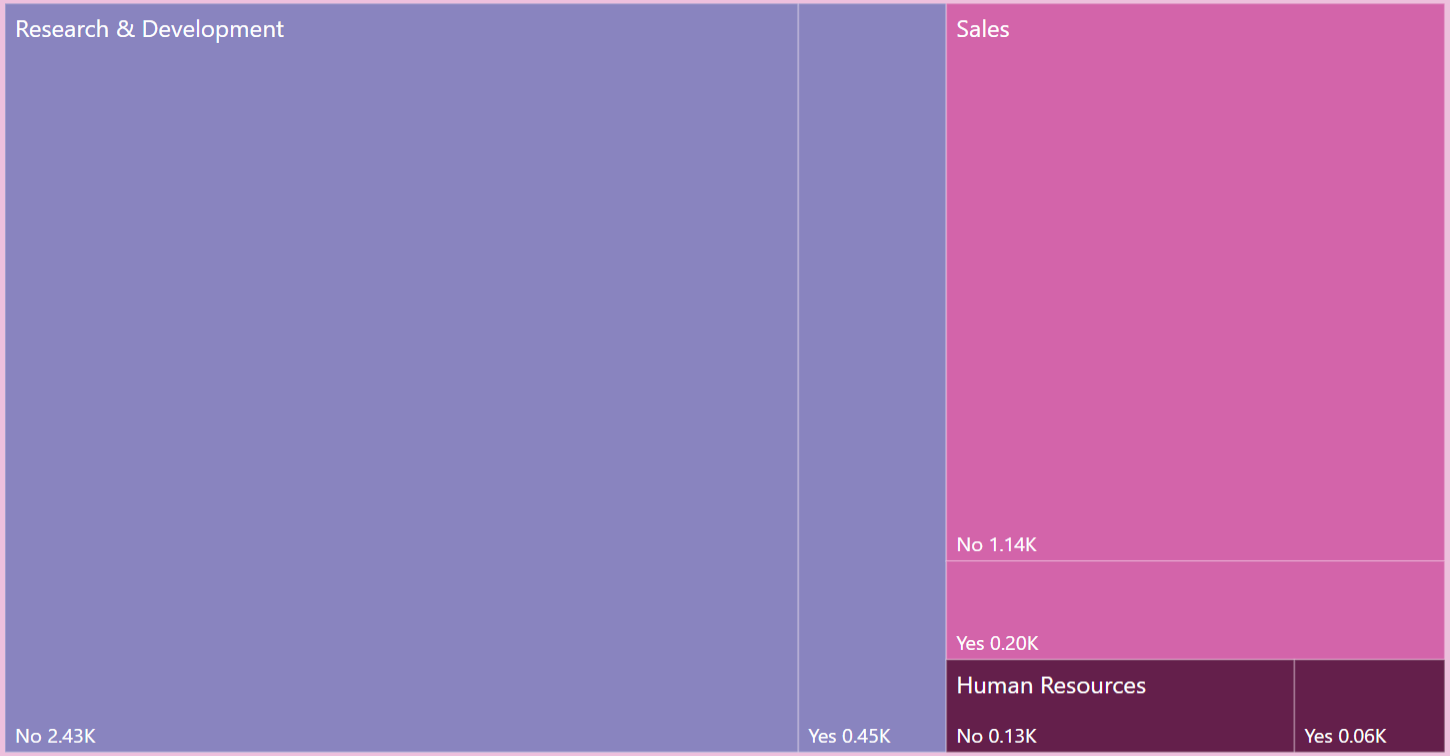

Count of Attrition by Age
60
40
20
0
20
30
40
50
60



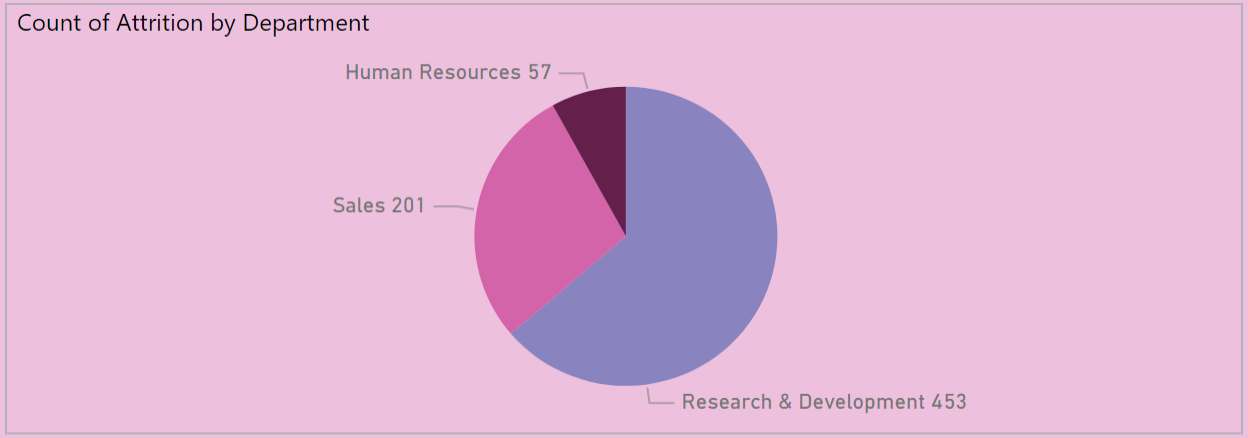


JobSatisfaction	Human Resources	Research & Development	Sales	**Total**
1	42	574	244	**860**
2	39	555	246	**840**
3	45	871	407	**1323**
4	63	868	436	**1367**
NA		15	5	**20**
Total	**189**	**2883**	**1338**	**4410**

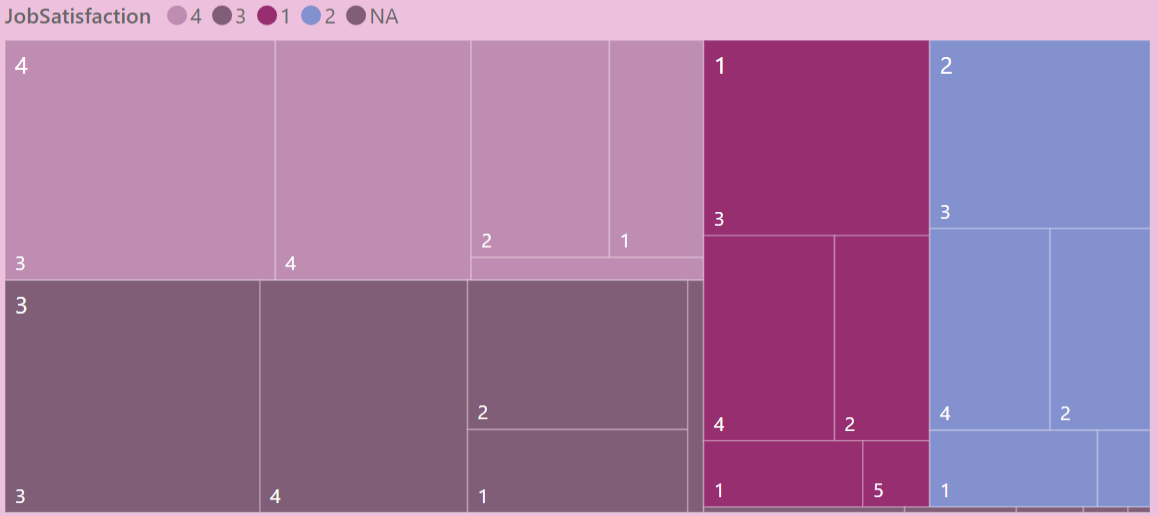
Count of Attrition by Department and Attrition

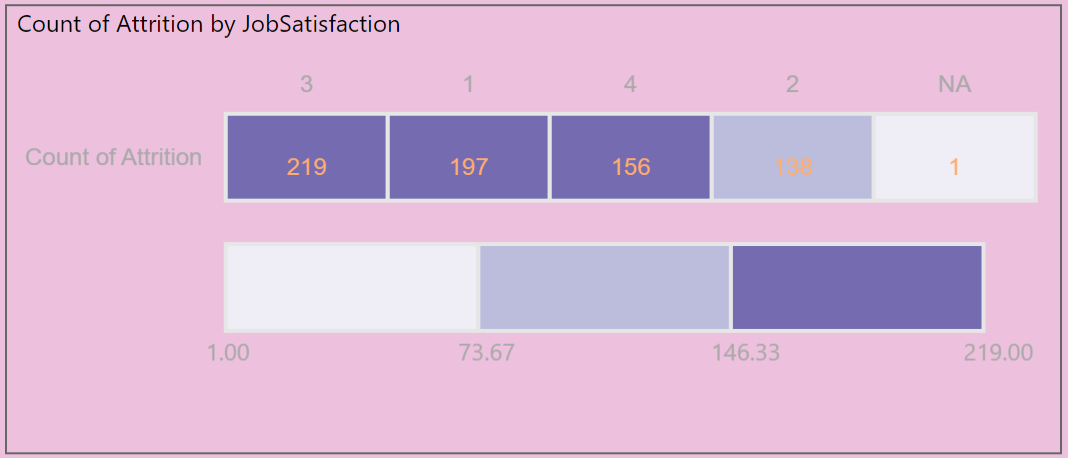


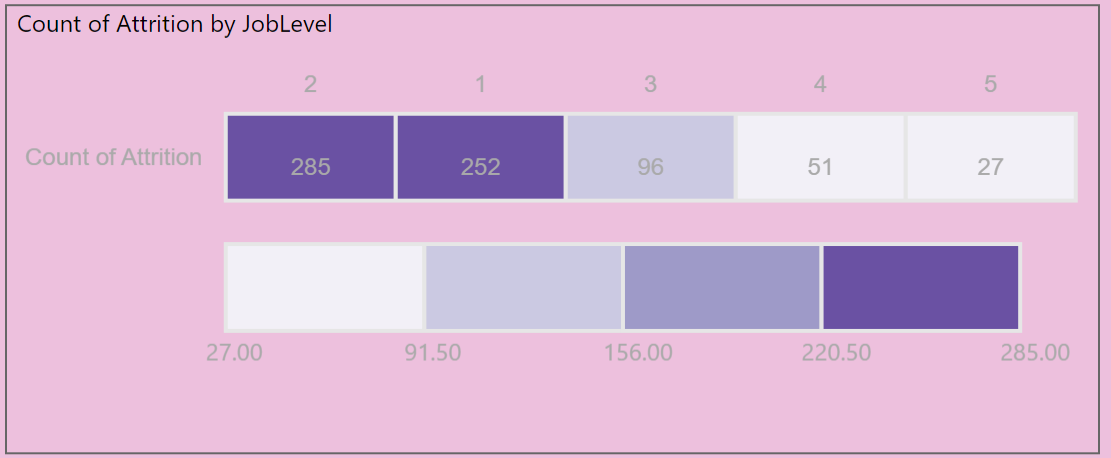
Count of Attrition by Department

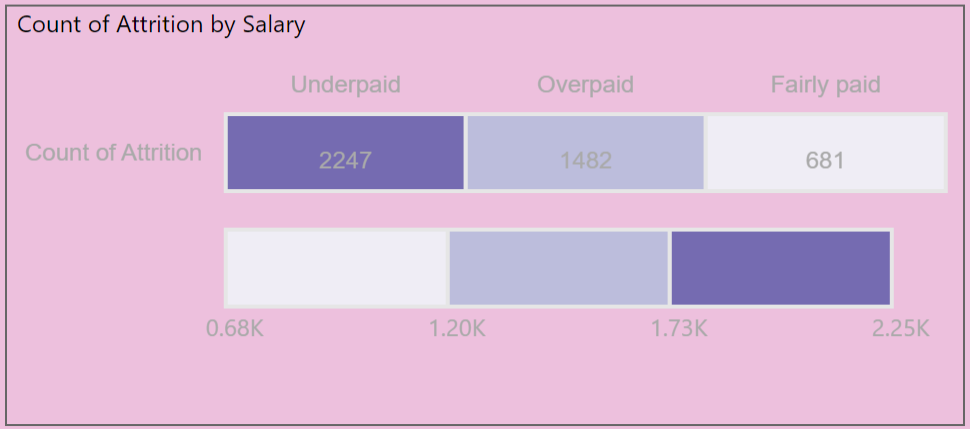


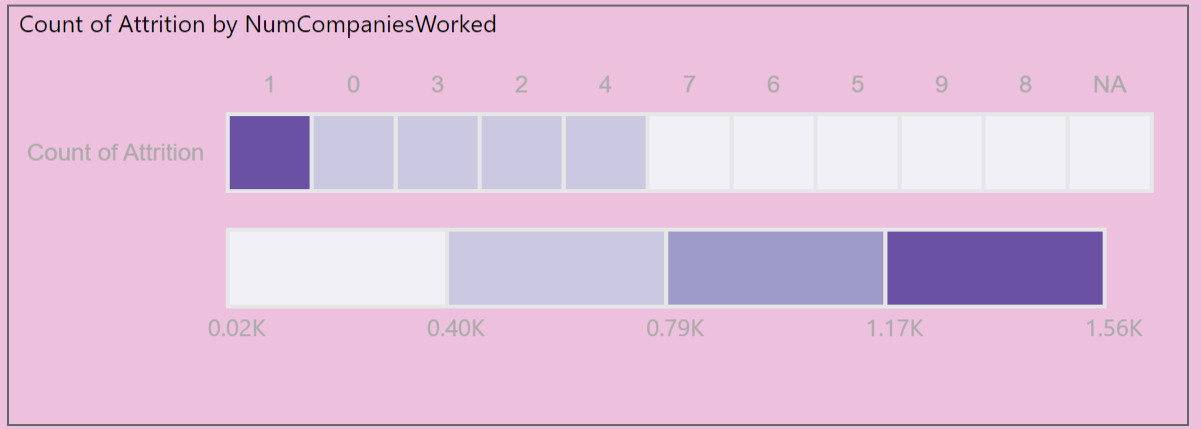
Count of Education by JobSatisfaction and Education

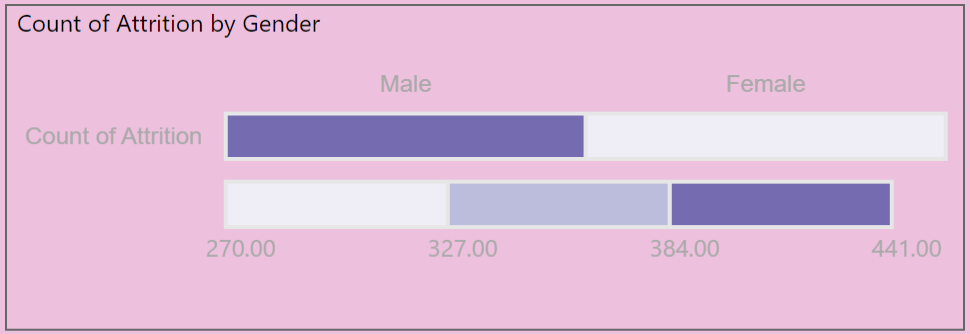


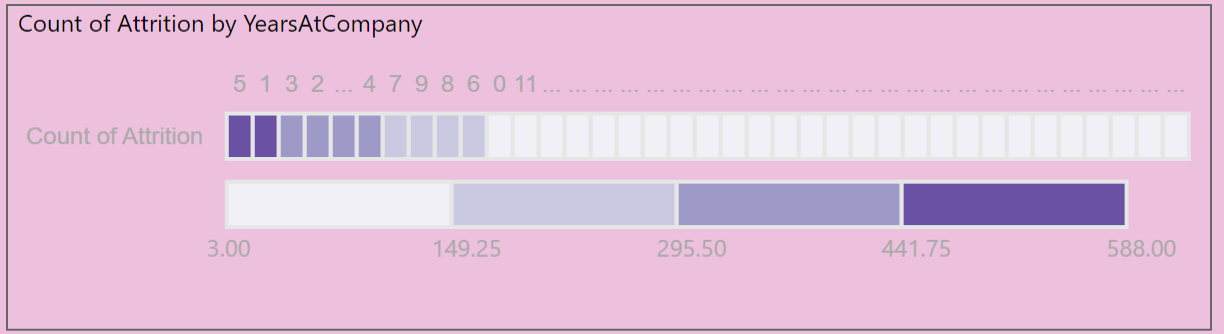

Count of Attrition by JobSatisfaction
3
1
4
2
NA
Count of Attrition
219
197
156
138
1
1.00
73.67
146.33
219.00


Count of Attrition by JobLevel
2
1
3
4
5
Count of Attrition
285
252
96
51
27
27.00
91.50
156.00
220.50
285.00

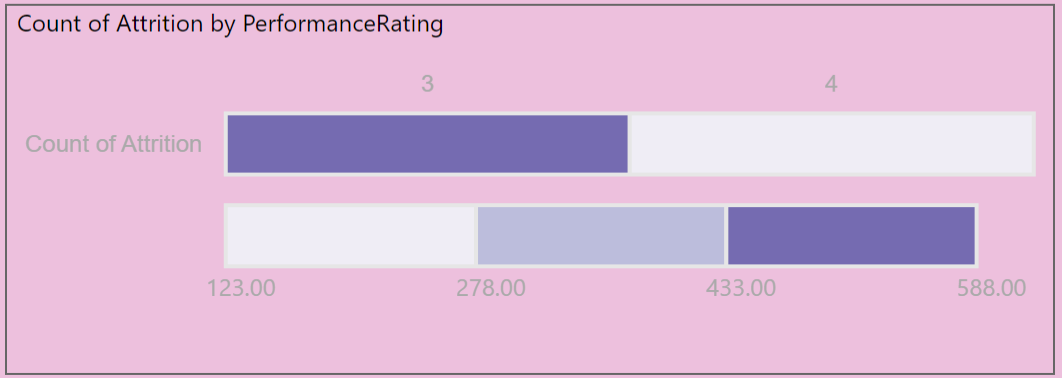

Count of Attrition by Salary
Underpaid
Overpaid
Fairly paid
Count of Attrition
2247
1482
681
0.68K
1.20K
1.73K
2.25K

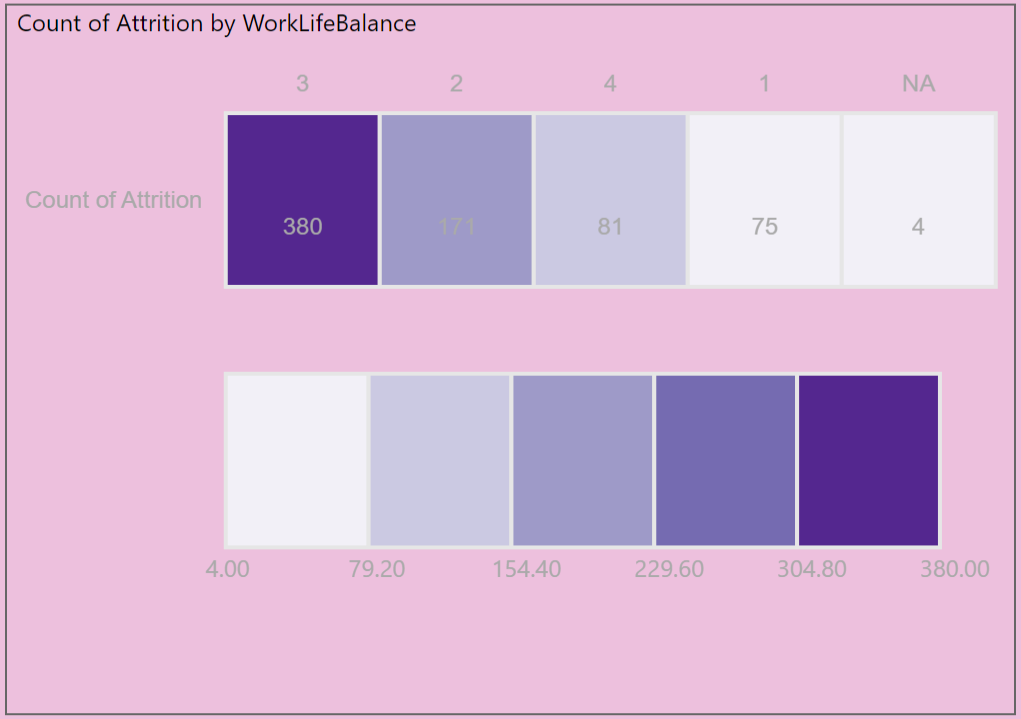

Count of Attrition by NumCompaniesWorked
1
0
3
2
4
7
6
5
9
8
NA
Count of Attrition
0.02K
0.40K
0.79K
1.17K
1.56K

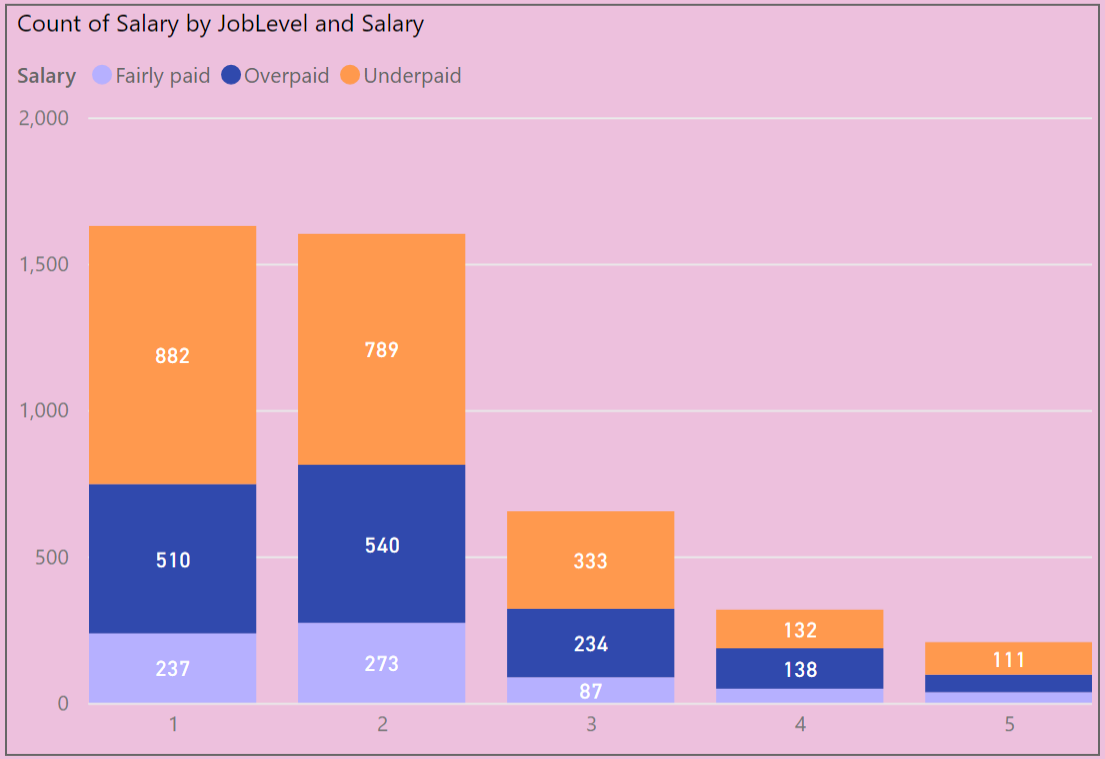

Count of Attrition by Gender
Male
Female
Count of Attrition
270.00
327.00
384.00
441.00

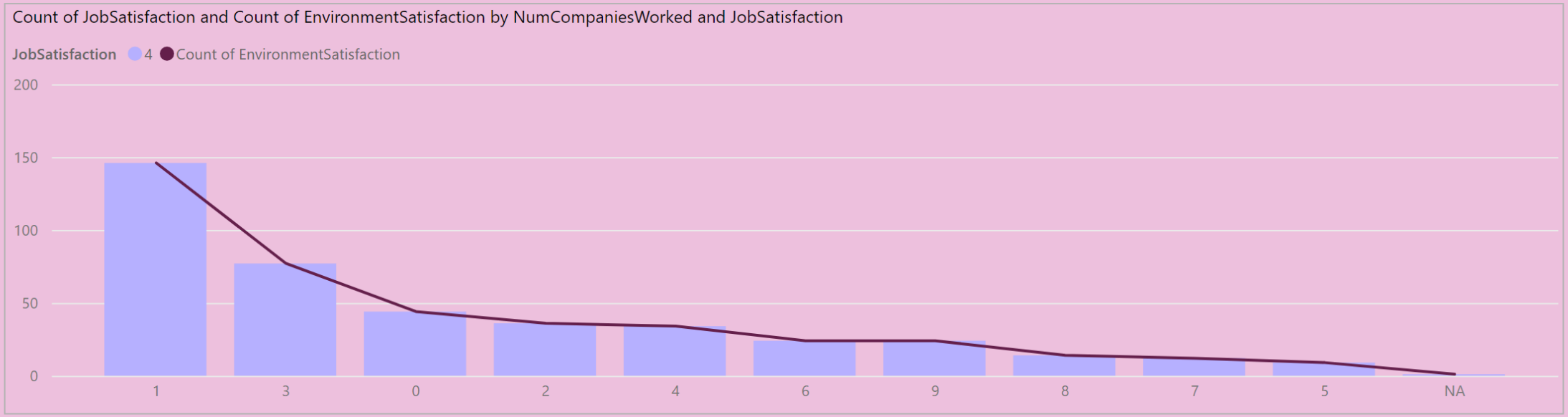

Count of Attrition by YearsAtCompany
5 1 3 2 ... 4 7 9 8 6 0 11
Count of Attrition
3.00
149.25
295.50
441.75
588.00

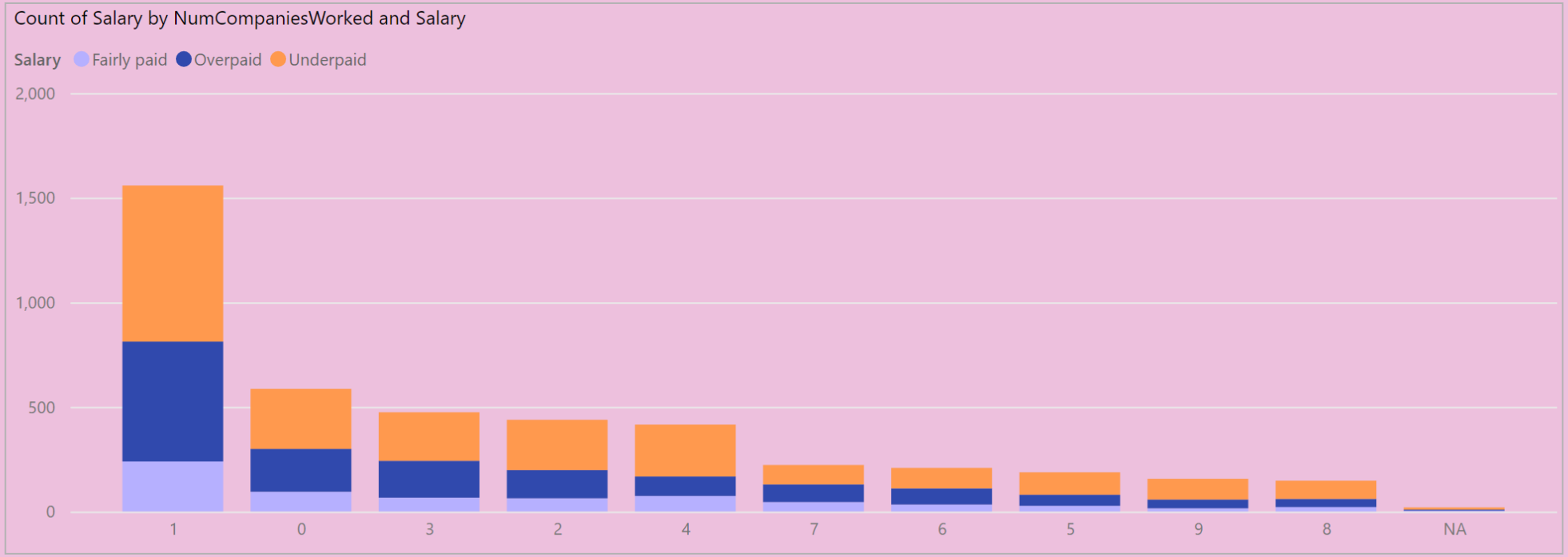

Count of Attrition by EnvironmentSatisfaction
1
3
4
2
NA
Count of Attrition
213
186
179
128
5
5.00
74.33
143.67
213.00

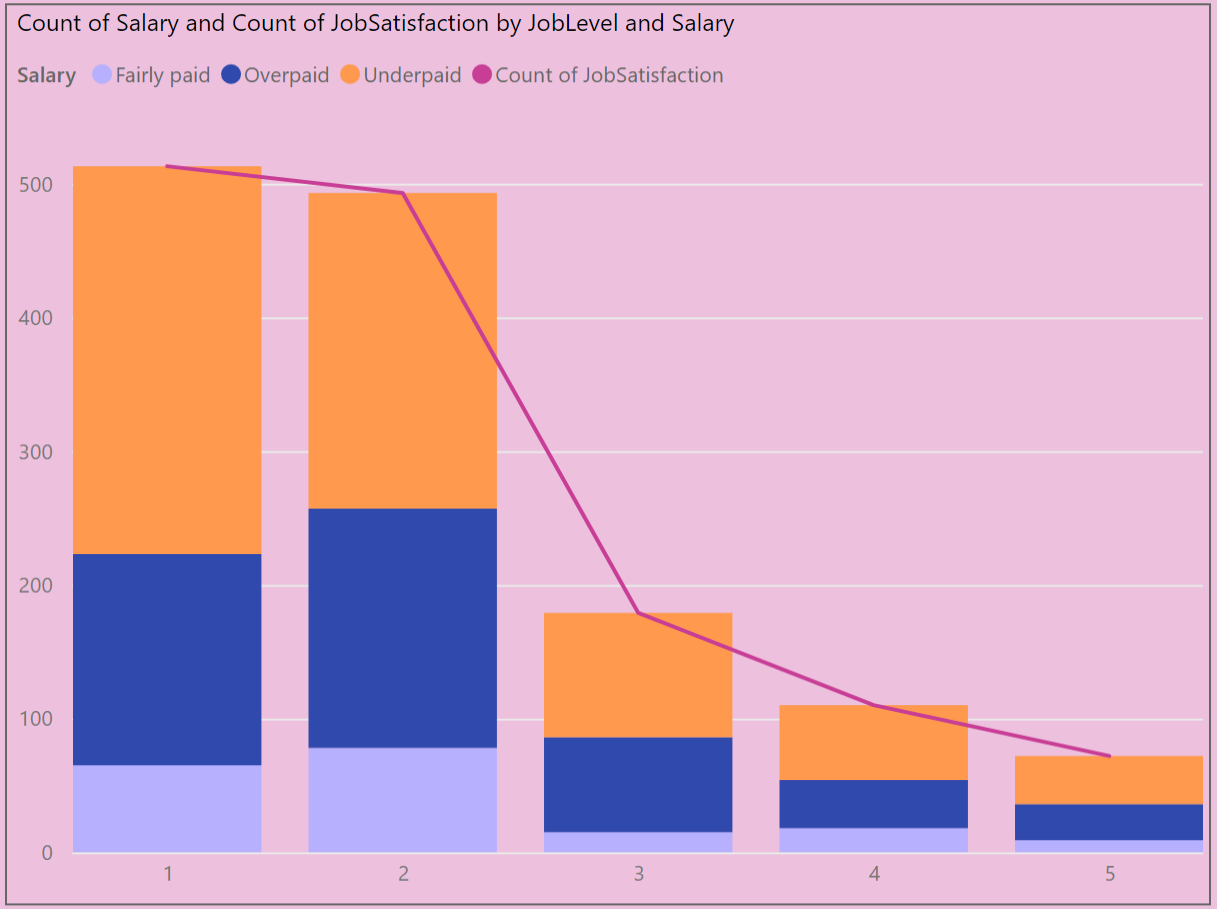

Count of Attrition by PerformanceRating
3
4
Count of Attrition
123.00
278.00
433.00
588.00

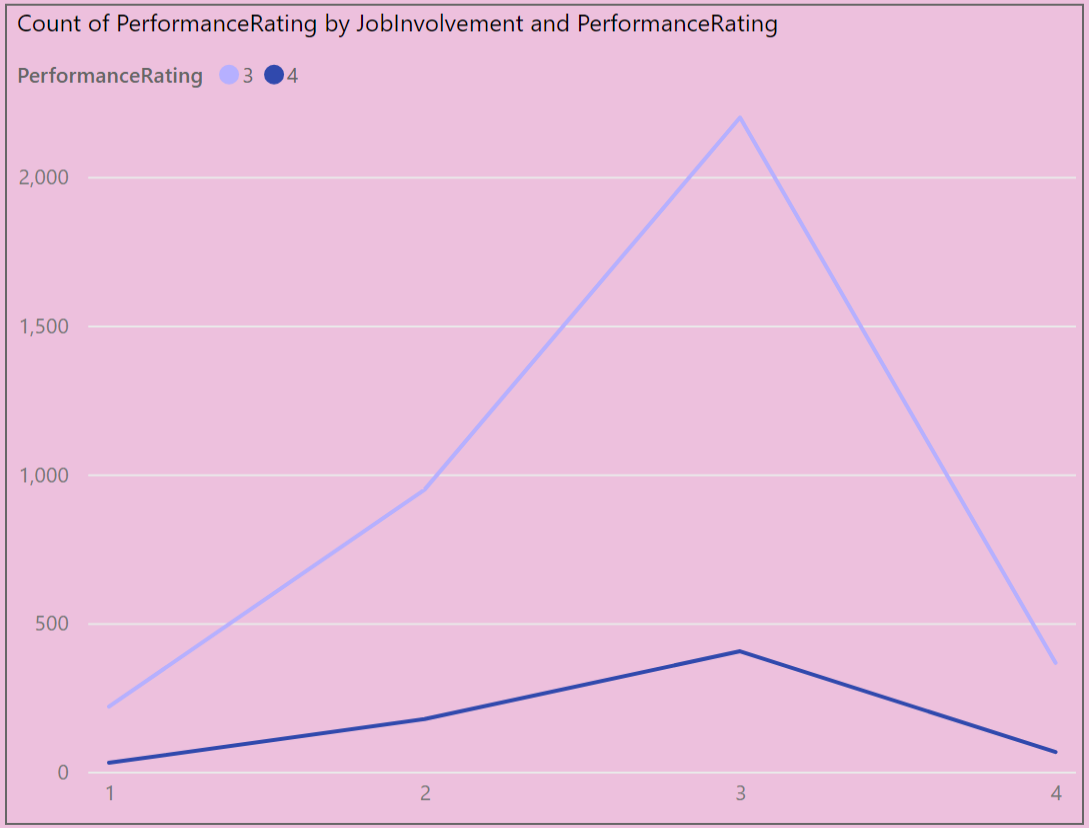

Count of Attrition by WorkLifeBalance
3
2
4
1
NA
Count of Attrition
380
171
81
75
4
4.00
79.20
154.40
229.60
304.80
380.00


Count of Salary by JobLevel and Salary
Salary
Fairly paid
Overpaid
Underpaid
2,000
1,500
1,000
500
0
882
789
510
540
333
234
132
138
111
237
273
87
1
2
3
4
5


Count of JobSatisfaction and Count of EnvironmentSatisfaction by NumCompaniesWorked and JobSatisfaction
JobSatisfaction
4
Count of EnvironmentSatisfaction
200
150
100
50
0
1
3
0
2
4
6
9
8
7
5
NA


Count of Salary by NumCompaniesWorked and Salary
Salary
Fairly paid
Overpaid
Underpaid
2,000
1,500
1,000
500
0
1
0
3
2
4
7
6
5
9
8
NA


Count of Salary and Count of JobSatisfaction by JobLevel and Salary
Salary
Fairly paid
Overpaid
Underpaid
Count of JobSatisfaction
500
400
300
200
100
0
1
2
3
4
5


Count of PerformanceRating by JobInvolvement and PerformanceRating
PerformanceRating
3
4
2,000
1,500
1,000
500
0
1
2
3
4

INTRODUCTION

1.1 ABOUT THE DATA

THE DATASET CONTAINS 29 COLUMNS AND 4411 ROWS CONTAINING INFORMATION ABOUT A COMPANY'S EMPLOYEE DETAILS - THE DEPARTMENT THEY WORK IN EDUCATION LEVEL, JOB ROLE, JOB AND ENVIRONMENT SATISFACTION AND WHETHER THEY HAVE A "YES" FOR ATTRITION.

1.2 OBJECTIVES

THE MAIN OBJECTIVES OF THIS PROJECT ARE TO:

(I) VISUALISE THE DATA TO SEE MEANINGFUL CORRELATIONS USING POWER BI.

(II) TO CALCULATE THE TOTAL NUMBER OF EMPLOYEES FOR AND AGAINST ATTRITION.

(III) TO ANALYSE THE ATTRITION RATES ACROSS VARIOUS DEPARTMENTS.

(IV) ANALYSING THE FACTORS THAT MIGHT CONTRIBUTE TO HIGH ATTRIBUTION RATES SUCH AS JOB SATISFACTION, MONTHLY INCOME, SALARY HIKES, ETC.

(V) BUILDING A DASHBOARD WHICH CAN HELP THE ORGANISATION IN MAKING DATA-DRIVEN DECISIONS.

1.3 DATA CLEANING

THERE ARE NO NULL VALUES IN THE DATA. WE REMOVE THE COLUMN EMPLOYEE COUNT AS IT IS IRRELEVANT IN THE ANALYSIS. EACH EMPLOYEEID REFERS TO A UNIQUE EMPLOYEE. THERE ARE NO NULL OR MISSING VALUES.

METHODOLOGY

* WE CALCULATE THE TOTAL NUMBER OF EMPLOYEES FOR THE ATTRITION VALUES “YES” AND “NO”. (CHART1)

* WE CREATE BAR GRAPHS TO DEDUCE WHICH AGE GROUPS ARE FOR AND AGAINST ATTRITION. (CHART 2 &3)

* A LINE GRAPH SHOWS THAT THE DEPARTMENTS LIKEWISE HAD THE SAME ORDER OF MOST SATISFIED AND LEAST SATISFIED EMPLOYEES IN THE ORDER MENTIONED ABOVE. (CHART 6). A PIVOT TABLE (CHART 5) DEPICTS THE NUMBER OF EMPLOYEES AND THEIR LEVEL OF JOB SATISFACTION IN EACH DEPARTMENT

* A LINE CHART (CHART 4) BETWEEN JOB SATISFACTION AND YEARS AT THE COMPANY SHOWS THAT ALTHOUGH THE NUMBER OF SATISFIED EMPLOYEES WAS MORE THAN THE NUMBER OF EMPLOYEES WHO WEREN’T IN THE BEGINNING YEARS, AS THE YEARS AT THE COMPANY PROGRESSED, THE SATISFACTION LEVELS DROPPED DOWN. EVENTUALLY, THE NUMBER OF MOST SATISFIED AND LEAST SATISFIED EMPLOYEES CONVERGED.

* A PIE CHART (CHART 8) IS USED FOR COMPARING ATTRITION BASED ON DEPARTMENT AND RESEARCH & DEVELOPMENT DEPARTMENT HAD THE HIGHEST ATTRITION ACCOUNTING FOR 453 EMPLOYEES FOLLOWED BY THE SALES DEPARTMENT ACCOUNTING FOR 201 EMPLOYEES AND THE HUMAN RESOURCES DEPARTMENT ACCOUNTING FOR 57 EMPLOYEES. A TREEMAP (CHART 7) ALSO DECODES THE NUMBER OF EMPLOYEES UNDER “YES” AND “NO” FOR ATTRITION IN EACH DEPARTMENT.

* A TREEMAP (CHART 9) BETWEEN JOB SATISFACTION AND EDUCATION LEVEL SHOWS THAT EMPLOYEES WITH HIGHER EDUCATION LEVELS WERE MORE SATISFIED THAN THOSE WITH LOWER EDUCATION LEVELS WITH THE COORDINATES OF (JOB SATISFACTION, EDUCATION LEVEL) NAMELY (4,3), (4,4). (3,3), (3,4) BEING GREATER IN PROPORTION.

* WE THEN CHECK THE CORRELATION BETWEEN VARIOUS FACTORS AND ATTRITION (CHART 10-17) TO DETERMINE THE CONTRIBUTING FACTORS FOR EMPLOYEE ATTRITION. FACTORS WHICH HIGH;Y INFLUENCED ATTRITION AMONG EMPLOYEES WERE:

JOB SATISFACTION, SALARY, JOB LEVEL, YEARS AT COMPANY, ENVIRONMENT SATISFACTION AND WORK LIFE BALANCE WHICH HAD A HIGH CORRELATION WITH ATTRITION, FOLLOWED BY NUMBER OF COMPANIES WORKED AT, GENDER AND PERFORMANCE RATING WHICH WERE SLIGHTLY LESS CORRELATED WITH ATTRITION IN COMPARISON.

* TO CATEGORISE THE SALARIES OF THE EMPLOYEES INTO HIGH, MEDIUM AND LOW, WE CALCULATE THE AVERAGE SALARY OF ALL THE EMPLOYEES.

AVERAGE SALARY FOR EDUCATION LEVEL 5 = 65,000
AVERAGE SALARY FOR EDUCATION LEVEL 5 IN RESEARCH AND DEVELOPMENT DEPARTMENT = 75,000
AVERAGE SALARY FOR EDUCATION LEVEL 5 IN HUMAN RESOURCES DEPARTMENT = 32,000
AVERAGE SALARY FOR EDUCATION LEVEL 5 IN SALES DEPARTMENT = 51,000

AVERAGE SALARY FOR EDUCATION LEVEL 1 = 61,000
AVERAGE SALARY FOR EDUCATION LEVEL 1 IN RESEARCH AND DEVELOPMENT DEPARTMENT = 63,000
AVERAGE SALARY FOR EDUCATION LEVEL 1 IN HUMAN RESOURCES DEPARTMENT = 52,000
AVERAGE SALARY FOR EDUCATION LEVEL 1 IN SALES DEPARTMENT = 58,000

THIS IMPLIES THAT IRRESPECTIVE OF THEIR EDUCATION LEVEL, HUMAN RESOURCES EMPLOYEES ARE HIGHLY UNDERPAID COMPARED TO OTHER DEPARTMENTS.

BASED ON THIS INFORMATION , WE WRITE A DAX SYNTAX TO CLASSIFY EMPLOYEE SALARIES INTO “FAIRLY PAID”, “UNDERPAID” AND “OVERPAID”.

* WE WRITE THE FOLLOWING DAX QUERY TO CLASSIFY EMPLOYEE SALARIES INTO “UNDERPAID”, “FAIRLY PAID” AND “UNDERPAID”.

```
Salary = IF(
    AND('Attrition data'[MonthlyIncome] > 50000, 'Attrition data'[MonthlyIncome] <= 65000),
    "Fairly paid",
    IF('Attrition data'[MonthlyIncome] < 50000,
        "Underpaid",
        "Overpaid"
    )
)
```

* THE NUMBER OF EMPLOYEES BEING UNDERPAID WAS HIGHER FOR LOWER JOB LEVELS. (CHART 19) HOWEVER, THE NUMBER OF EMPLOYEES BEING UNDERPAID WAS HIGHER THAN THOSE BEING PAID FAIRLY AND OVERPAID, IRRESPECTIVE OF THE JOB LEVEL. THESE NUMBERS DECREASE WITH AN INCREASE IN JOB LEVEL.

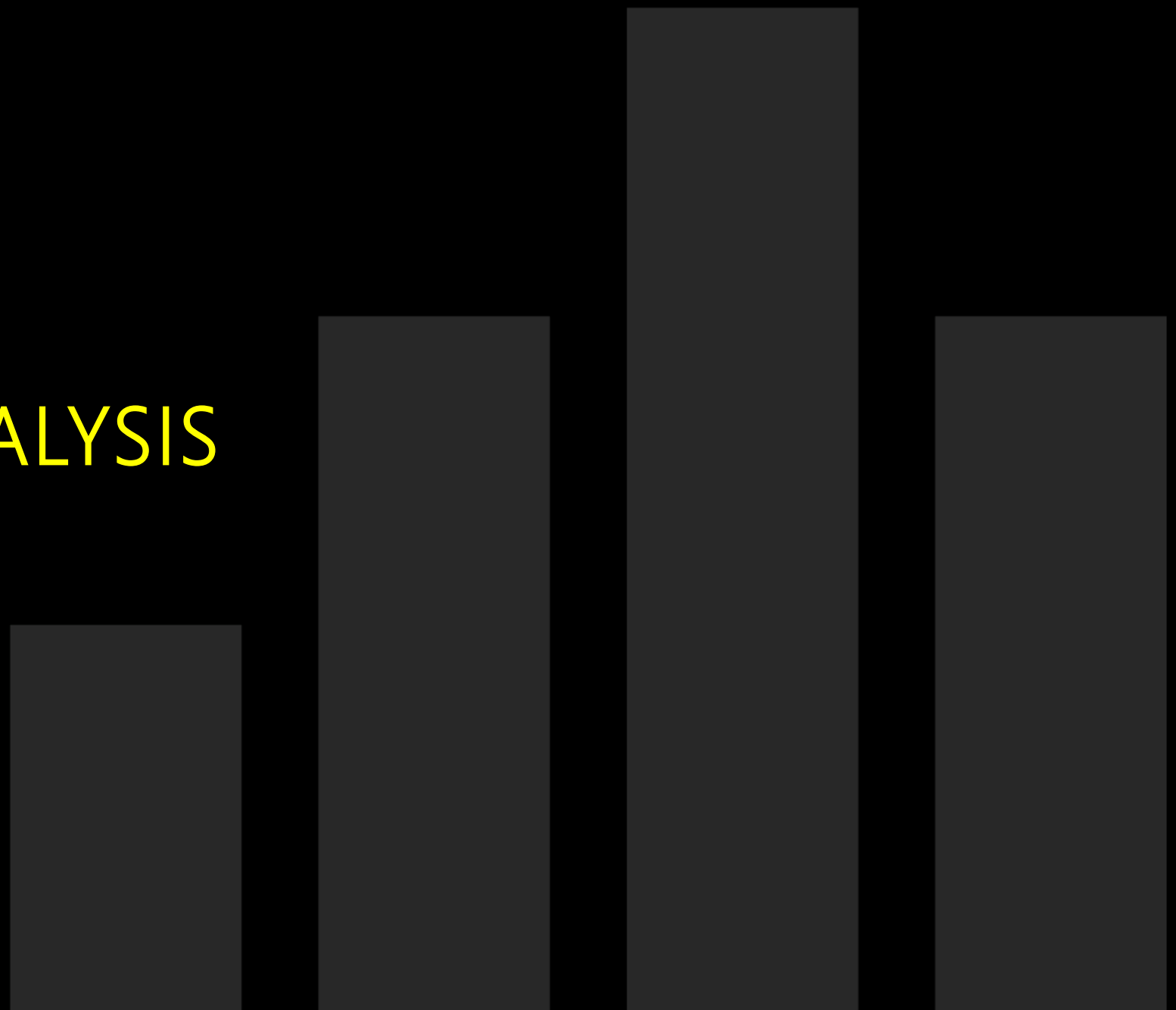
* THE NUMBER OF EMPLOYEES SATISFIED WITH THE JOB DECREASED WITH AN INCREASING NUMBER OF COMPANIES THEY WORKED IN. (CHART 20) EMPLOYEES WHO WORKED IN 1 COMPANY WERE MORE SATISFIED WITH THEIR JOB AND ALSO WITH THEIR ENVIRONMENT THAN THOSE WHO WORKED IN 6 COMPANIES AND SO ON.

* A BAR GRAPH WAS PLOTTED AGAINST THE NUMBER OF COMPANIES WORKED AND SALARY (CHART 21). NUMBER OF EMPLOYEES WHO WORKED IN ONLY 1 COMPANY WHO WERE UNDERPAID WERE SIGNIFICANTLY HIGHER THAN THOSE WHO WORKED IN 0, 2,3 OR 4 COMPANIES.

* ALTHOUGH NUMBER OF EMPLOYEES WHO WORKED IN 1 COMPANY WAS HIGHER THAN THOSE WHO WORKED IN 2 COMPANIES, THE PROPORTION OF OVERPAID EMPLOYEES WAS HIGHER AMONG EMPLOYEES WHO WORKED IN 2 COMPANIES. THE PROPORTION OF UNDERPAID EMPLOYEES WAS HIGHER AMONG EMPLOYEES WHO WORKED IN 1 COMPANY, YET COUNT OF JOB SATISFACTION WAS HIGHER AMONG EMPLOYEES WHO WORKED IN 1 COMPANY AND IT DECREASED AS THE NUMBER OF COMPANIES WORKED AT INCREASED. (CHART 22)

* JOB INVOLVEMENT AND PERFORMANCE RATING HAD A SOMEWHAT LINEAR RELATIONSHIP WHICH IMPLIES THAT PERFORMANCE RATING INCREASES WITH AN INCREASE IN JOB INVOLVEMENT AND DECREASES WITH A DECREASE IN JOB INVOLVEMENT. (CHART 23)

IMDB MOVIE ANALYSIS



INTRODUCTION

1.1 ABOUT THE PROJECT

THE DATASET USED IS RELATED TO IMDB MOVIES FROM KAGGLE. THERE ARE 28 COLUMNS AND 5043 ROWS IN THE DATASET.

1.2 DATA CLEANING

THIS STEP INVOLVES PREPROCESSING THE DATA TO MAKE IT SUITABLE FOR ANALYSIS. IT INCLUDES HANDLING MISSING VALUES, REMOVING DUPLICATES, CONVERTING DATA TYPES IF NECESSARY, AND POSSIBLY FEATURE ENGINEERING.

WE HAVE FEW COLUMNS THAT ARE IRRELEVANT TO OUR ANALYSIS SUCH AS ACTOR NAME,ACTOR LIKES, COLOR, DIRECTOR_FACEBOOK_LIKES, FACENUMBER_IN_POSTER, MOVIE_IMDB_LINK, ASPECT_RATIO, CONTENT_RATING, NUM_VOTED_USERS, MOVIE_FACEBOOK_LIKES HENCE WE DELETE THOSE ROWS. WE PLACE NECESSSARY COLUMNS NEXT TO EACH OTHER TO MAKE THE ANALYSIS EASY.

WE CAN FIND EMPTY CELLS USING THE FIND&SELECT OPTION IN EXCEL AND DELETE THEIR RESPECTIVE ROWS AS PART OF DATA CLEANING BUT DELETING ALL ROWS TOGETHER WOULD LEAD TO DELETING A HUGE PART OF THE DATA. TO AVOID DATA LOSS, WE CAN DELETE ROWS FOR EACH SECTION OF THE ANALYSIS. FOR INSTANCE, FOR ANALYSIS DIFFERENT GENRES AND THEIR IMDB SCORE, WE CAN FILTER OUT ROWS OF GENRE AND SCORE IN A SEPARATE SHEET AND ANALYSE THEM TO PREVENT DELETING DATA FROM IMPORTANT COLUMNS LIKE BUDGET, LANGUAGE, YEAR THAT MIGHT NOT BE EMPTY.

INTRODUCTION

1.3 REMOVING DUPLICATES

DUPLICATE ROWS CAN BE R3EMOVED BY USING THE DATA TAB AND SELECTING THE “REMOVE DUPLICATES” OPTION. 121 ROWS ARE DELETED WHICH WERE DETECTED TO BE DUPLICATES.

1.4 FEATURE ENGINEERING

FEATURE ENGINEERING IS THE PROCESS OF CREATING NEW VARIABLES (FEATURES) OR MODIFYING EXISTING ONES TO IMPROVE ANALYSIS AND MODEL PERFORMANCE. IT HELPS IN UNCOVERING HIDDEN PATTERNS IN THE DATA. IN OUR DATASET, WE CAN USE FEATURE ENGINEERING ON COLUMNS SUCH AS SPLITTLING GENRE, YEAR (TO CLASSIFY YEARS INTO DECADES). CREATING A PROFIT COLUMN WHERE PROFIT = GROSS - BUDGET AND CONVERTING IMDB SCORE INTO CATEGORIES LIKE "LOW", "AVERAGE", "HIGH". WE CAN DO THIS DURING INDICIDUAL ANALYSIS LIKE FOR REMOVING NULL VALUES.

A. MOVIE GENRE ANALYSIS:
ANALYZE THE DISTRIBUTION OF MOVIE GENRES AND THEIR IMPACT ON THE IMDB SCORE

ON A NEW EXCEL SHEET, WE TAKE THE COLUMNS - MOVIE NAME, GENRE AND IMDB SCORE AND REMOVE NULL VALUES AND DUPLICATE ROWS. WE PERFORM FEATURE ENGINEERING TO EXTRACT THE DIFFERENT GENRES.

1. SPLIT GENRES INTO SEPARATE COLUMNS (TEXT TO COLUMNS METHOD)
SELECT THE "GENRES" COLUMN.
GO TO DATA → CLICK TEXT TO COLUMNS.
CHOOSE DELIMITED → CLICK NEXT.
SELECT OTHER AND ENTER | (PIPE SYMBOL) → CLICK FINISH.

2. CREATE BINARY COLUMNS FOR EACH GENRE (ONE-HOT ENCODING)
CREATE NEW COLUMN HEADERS FOR EACH GENRE (E.G., ACTION, COMEDY, DRAMA).
=IF(COUNTIF($D2:$L2,"ACTION")>0,1,0) , REPEAT FOR EACH GENRE

3. DATA ANALYSIS
FIND MEAN, MEDIAN IMDB SCORE FOR EACH GENRE

GENRE	AVERAGE IMDB SCORE	MEDIAN IMDB SCORE
Action	6.2	6.3
Adventure	6.4	6.6
Fantasy	6.3	6.4
Sci-Fi	6.3	6.4
Thriller	6.3	6.4
Romance	6.4	6.5
Animation	6.6	6.7
Family	6.2	6.4
Musical	6.5	6.7
Drama	6.8	6.9
Crime	6.6	6.6
Western	6.7	6.8
Mystery	6.5	6.6
Horror	5.8	6.6
Biography	7.1	5.9
War	7.1	7.2
History	7.1	7.2
Sport	6.6	6.8
Documentary	7.2	7.4

	Action	Adventure	Fantasy	Sci-Fi	Thriller	Romance	Animation	Family	Musical	Drama	Crime	Western	Mystery	Horror	Biography	W
SCORE	6.6	6.7	6.7	6.7	6.4	6.5	6.7	6.7	7	6.7	6.6	6.5	6.6	6.2	7	
MDB Sc	1.3	1.3	1.4	1.5	1.1	1.0	1.3	1.5	1.5	0.9	1.1	1.1	1.2	1.3	0.5	
Nation	1.1	1.1	1.2	1.2	1.1	1.0	1.1	1.2	1.2	1.0	1.0	1.1	1.1	1.1	0.7	

B. MOVIE DURATION ANALYSIS

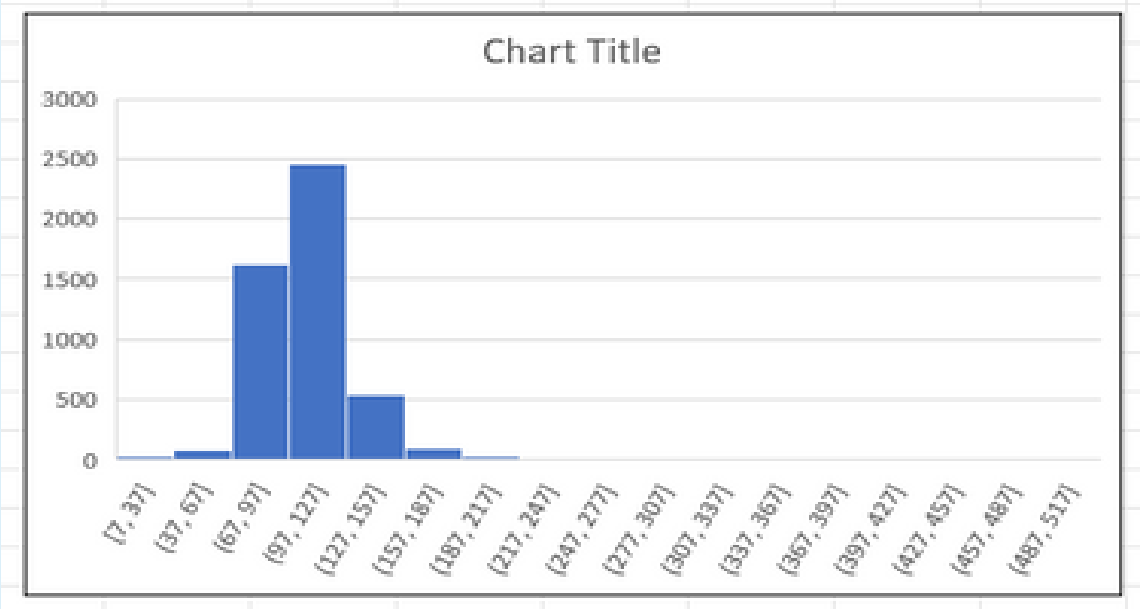
ANALYZE THE DISTRIBUTION OF MOVIE DURATIONS AND ITS IMPACT ON THE IMDB SCORE.

ON A NEW EXCEL SHEET, WE TAKE THE COLUMNS - MOVIE NAME, DURATION AND IMDB SCORE AND REMOVE NULL VALUES AND DUPLICATE ROWS.

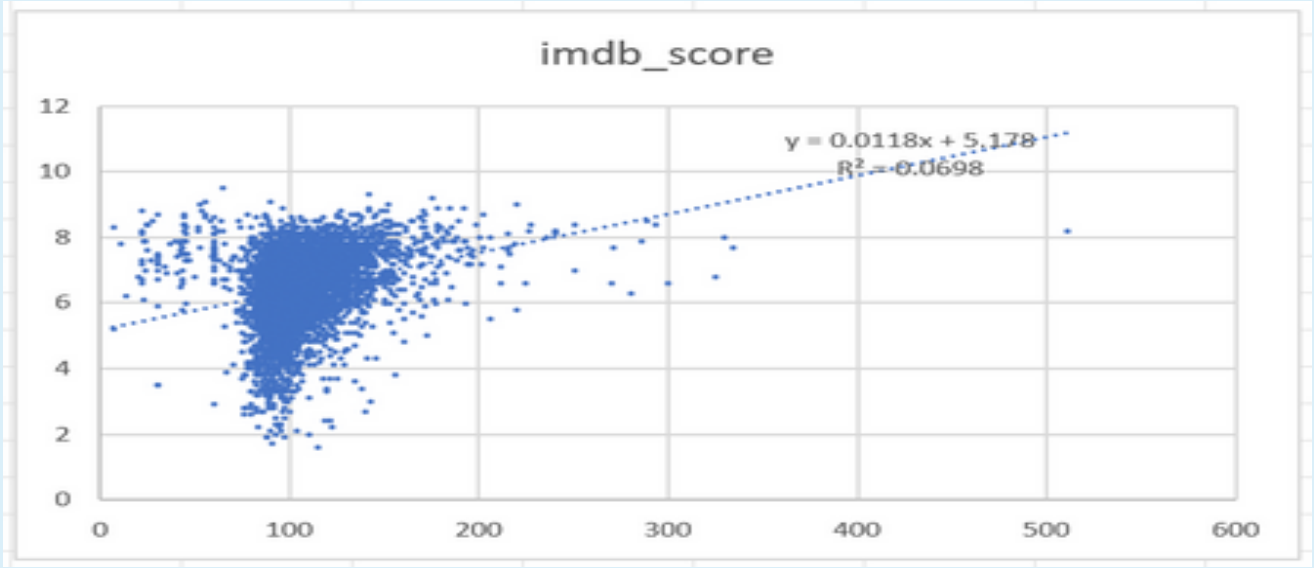
WE CREATE A HISTOGRAM FOR THE MOVIE DURATIONS AND INCREASE THE BIN SIZE TO SEE THE DISTRIBUTION OF DURATIONS. MOST MOVIES HAVE A DURATION BETWEEN 80-150 MINUTES

WE CALCULATE THE MEAN, MEDIAN, MODE, AND RANGE OF THE DURATION FOR ALL MOVIES. WE ALSO CALCULATE THE CORRELATION BETWEEN DURATION AND IMDB SCORE. IT IS 0.2642.
SINCE THE VALUE IS GREATER THAN 0, IT SUGGESTS THAT LONGER MOVIES TEND TO HAVE SLIGHTLY HIGHER IMDB SCORES.
HOWEVER, BECAUSE 0.2642 IS CLOSER TO 0 THAN TO 1, THE RELATIONSHIP IS WEAK, MEANING MOVIE DURATION ALONE IS NOT A STRONG PREDICTOR OF IMDB RATINGS. HENCE, LONGER MOVIES MIGHT GET HIGHER RATINGS, BUT THE EFFECT IS SMALL.

WE THEN CREATE A SCATTER PLOT BETWEEN MOVIE DURATION AND IMDB SCORES ANDINSERT A TRENDLINE. THE R^2 VALUE = 0.0698, MEANING ONLY 6.98% OF THE VARIATION IN IMDB SCORES CAN BE EXPLAINED BY MOVIE DURATION. THIS IS A VERY WEAK RELATIONSHIP, MEANING THAT DURATION ALONE IS NOT A STRONG PREDICTOR OF IMDB SCORE.



AVERAGE	107.099
MEDIAN	103
MODE	90
MAXIMUM	511
MINIMUM	7
RANGE	504
STANDARD DEVIATION	25.2798
CORRELATION	0.26424



C. LANGUAGE ANALYSIS SITUATION
EXAMINE THE DISTRIBUTION OF MOVIES BASED ON THEIR LANGUAGE.

ON A NEW EXCEL SHEET, WE TAKE THE COLUMNS - MOVIE NAME, LANGUAGE AND IMDB SCORE AND REMOVE NULL VALUES AND DUPLICATE ROWS.BY USING A PIVOT COLUMN, WE FIND THE TOP 5 LANGUAGES BY COUNT OF MOVIESPER EACH LANGUAGE.

AFTER WE FIND THE TOP 5 LANGUAGES. WE FIND THE AVERAGE (MEAN), MEDIAN AND STNADARD DEVIATION OF IMDB SCORE FOR EACH LANGUAGE USING THE CONDITION:
=AVERAGEIF(B:B, "ENGLISH", C:C),
=MEDIAN(IF(B:B="FRENCH", C:C)),
=STDEV.P(IF(B:B="ENGLISH", C:C))

WE REPEAT THIS FOR THE OTHER 4 LANGUAGES TO FIND THEIR IMPACT ON THE IMDB SCORES.

SINCE WE HAVE TO USE PERCENTILE FUNCTION TO IDENTIFY DIRECTORS WITH HIGHEST SCORES, WE USE THE FUNCTION PERCENTILE.INC(C:C, 0.9) WHERE C IS THE IMDB SCORE COLUMN. THE RESULT IS 7.7 WHICH MEANS THAT 90% OF MOVIES HAVE AN IMDB SCORE BELOW 8.2, AND ONLY 10% OF MOVIES HAVE A SCORE ABOVE THIS.THIS HELPS US IDENTIFY TOP DIRECTORS—THOSE WHOSE AVERAGE IMDB SCORE EXCEEDS THE 90TH PERCENTILE ARE IN THE TOP 10% OF DIRECTORS.
TO FUTHER FILTER OUT THE TOP 10%, WE GIVE A CONDITIONAL STATEMENT TO CLASSIFY THE DIRECTOR AS TOP 10% OR BELOW 90TH PERCENTILE, I.E, IF(AVERAGEIF(B:B, B2, C:C) >= 7.7, "TOP 10%", "BELOW 90TH PERCENTILE"). INCASE WE WANT THE TOP 10 DIRECTORS OUT OF THIS TOP 10%, WE ALREADY HAVE THE RESULT ABOVE THAT WE OBTAINED USING A PIVOT TABLE.

MEAN IMDB SCORE FOR TOP 5 LANGUAGES	
English	6.39365
French	7.03836
Spanish	6.9375
Hindi	6.63214
Mandarin	6.7875

Row Labels	Count of movie_title
English	4584
French	73
Hindi	28
Mandarin	24
Spanish	40
Grand Total	**4749**

	COUNT	MEAN	MEDIAN	STANDARD DEVIATION
English	4584	6.39365	7.2	1.13
French	73	7.03836	7.15	0.72
Spanish	40	6.9375	6.95	0.84
Hindi	28	6.63214	7.05	1.37
Mandarin	24	6.7875	6.5	1.02

D. DIRECTOR ANALYSIS
INFLUENCE OF DIRECTORS ON MOVIE RATINGS.

REMOVE NULL VALUES, AND DUPLICATE ROWS.

BY USING A PIVOT COLUMN, WE FIND THE AVERAGE IMDB SCORE FOR EACH DIRECTOR BY PLACING THE DIRECTOR COLUMN IN ROWS AND AVERAGE OF IMDB SCORE AS VALUES. WE OBTAIN THE TOP 10 DIRECTORS BY AVERAGE OF IMDB SCORE BY USING SORT AND FILTER.

SINCE WE HAVE TO USE PERCENTILE FUNCTION TO IDENTIFY DIRECTORS WITH HIGHEST SCORES, WE USE THE FUNCTION PERCENTILE.INC(C:C, 0.9) WHERE C IS THE IMDB SCORE COLUMN. THE RESULT IS 7.7 WHICH MEANS THAT 90% OF MOVIES HAVE AN IMDB SCORE BELOW 8.2, AND ONLY 10% OF MOVIES HAVE A SCORE ABOVE THIS.THIS HELPS US IDENTIFY TOP DIRECTORS—THOSE WHOSE AVERAGE IMDB SCORE EXCEEDS THE 90TH PERCENTILE ARE IN THE TOP 10% OF DIRECTORS.
TO FUTHER FILTER OUT THE TOP 10%, WE GIVE A CONDITIONAL STATEMENT TO CLASSIFY THE DIRECTOR AS TOP 10% OR BELOW 90TH PERCENTILE, I.E, IF(AVERAGEIF(B:B, B2, C:C) >= 7.7, "TOP 10%", "BELOW 90TH PERCENTILE"). INCASE WE WANT THE TOP 10 DIRECTORS OUT OF THIS TOP 10%, WE ALREADY HAVE THE RESULT ABOVE THAT WE OBTAINED USING A PIVOT TABLE.

Row Labels	Average of imdb_score
John Blanchard	9.5
Sadyk Sher-Niyaz	8.7
Mitchell Altieri	8.7
Cary Bell	8.7
Mike Mayhall	8.6
Charles Chaplin	8.6
Damien Chazelle	8.5
Ron Fricke	8.5
Raja Menon	8.5
Majid Majidi	8.5
Grand Total	**8.68**

	A	B	C	D
1	movie_title	director_name	imdb_score	Percentile
2	AvatarÂ	James Cameron	7.9	Top 10%
5	The Dark Knight RisesÂ	Christopher Nolan	8.5	Top 10%
6	John CarterÂ	Andrew Stanton	6.6	Top 10%
8	TangledÂ	Nathan Greno	7.8	Top 10%
9	Avengers: Age of UltronÂ	Joss Whedon	7.5	Top 10%
18	The AvengersÂ	Joss Whedon	8.1	Top 10%
21	The Hobbit: The Battle of the Five ArmiesÂ	Peter Jackson	7.5	Top 10%
24	The Hobbit: The Desolation of SmaugÂ	Peter Jackson	7.9	Top 10%
26	King KongÂ	Peter Jackson	7.2	Top 10%
27	TitanicÂ	James Cameron	7.7	Top 10%
44	Toy Story 3Â	Lee Unkrich	8.3	Top 10%
58	WALLÂ·EÂ	Andrew Stanton	8.4	Top 10%
66	The Dark KnightÂ	Christopher Nolan	9	Top 10%
67	UpÂ	Pete Docter	8.3	Top 10%
78	Inside OutÂ	Pete Docter	8.3	Top 10%
89	Big Hero 6Â	Don Hall	7.9	Top 10%
90	Wreck-It RalphÂ	Rich Moore	7.8	Top 10%
93	How to Train Your DragonÂ	Dean DeBlois	8.2	Top 10%
96	InterstellarÂ	Christopher Nolan	8.6	Top 10%
97	InceptionÂ	Christopher Nolan	8.8	Top 10%

E. BUDGET ANALYSIS
EXPLORE THE RELATIONSHIP BETWEEN MOVIE BUDGETS AND THEIR FINANCIAL SUCCESS.

ON A NEW EXCEL SHEET, WE TAKE THE COLUMNS - MOVIE NAME, BUDGET, GROSS AND IMDB SCORE- REMOVE NULL VALUES, AND DUPLICATE ROWS.

WE CALCULATE THE PROFIT FOR EACH MOVIE BY CALCULATING THE DIFFERENCE BETWEEN THE BUDGET AND GROSS, I.E., PROFIT = GROSS - BUDGET
AFTER CALCULATING PROFIT FOR EACH MOVIE, WE FIND THE MOVIE WITH THE HIGHEST PROFIT BY CALCULATING THE MAXIMUM PROFIT AND INDEXING IT TO ITS RESPECTIVE MOVIE.

HIGHEST PROFIT = MAX(E:E)
MOVIE WITH HIGHEST PROFIT = INDEX (A:A, MATCH(MAX(E:E), E:E,0))
WHERE A - MOVIE NAME AND E - PROFIT

THE MOVIE WITH THE HIGHEST PROFIT IS AVATAR WITH A PROFIT OF 523505847.

TO ANALYSE THE CORRELATION BETWEEN MOVIE BUDGETS AND GROSS EARNINGS, WE USE THE CORREL FUNCTION IN EXCEL.

THE CORRELATION COEFFICIENT 0.0966 IS VERY CLOSE TO 0, INDICATING A WEAK POSITIVE RELATIONSHIP BETWEEN MOVIE BUDGET AND GROSS EARNINGS.

THIS MEANS THAT :

- BUDGET DOES NOT STRONGLY DETERMINE FINANCIAL SUCCESS.
- SOME LOW-BUDGET MOVIES MAY EARN A LOT, WHILE SOME HIGH-BUDGET MOVIES MIGHT UNDERPERFORM.
- OTHER FACTORS (E.G., GENRE, DIRECTOR, MARKETING, REVIEWS) LIKELY HAVE A BIGGER IMPACT ON EARNINGS.

correlation between budget and gross	0.096619736
highest profit	523505847
movie with highest profit	AvatarÂ