Title: People data analysis in a financial institution

0. Executive Summary

This report presents an analysis of HR data from one of financial institutions in the US, examining the impact of remote work, certifications, and education on employee performance and salaries. Additionally, it analyzes feedback from intern and manager surveys to assess sentiments and identify areas for improvement. Two datasets were analyzed: 1) HR data including headcount history, action history, and education details from 2016-2024, 2) Intern surveys from 2018-2021, including feedback from interns and managers.

Our analysis aimed to test the following hypotheses: 1) Remote work negatively impacts performance ratings, 2) More certifications lead to higher annual salaries, 3) Advanced degrees result in better performance ratings, 4) Identifying common keywords and sentiment in intern and manager feedback. For this, we used data analytics skills such as numerous regression models and NLP techniques.

Key findings from our analysis include:

Remote Work and Performance Ratings: Remote work is associated with an 18.3% decrease in the likelihood of higher performance ratings. However, this effect is mitigated by higher salaries or positions.

Certifications and Salaries: Employees without certifications earn an average annual salary of $110,127. Each additional certification is associated with an average salary increase of $4,372.

Education and Performance Ratings: Advanced degrees do not consistently predict higher performance ratings, suggesting that other factors influence performance.

Intern and Manager Feedback: Sentiment analysis revealed that virtual work negatively impacted performance and engagement. Interns and managers noted challenges related to technical issues and a lack of in-person interaction.

Based on these findings, we suggest the following recommendations:

Professional Development: Encourage certifications to enhance skills and productivity, integrating them with career progression paths.

Remote Work Policies: Reevaluate remote work policies through A/B testing different work arrangements to optimize performance and collaboration.

Event Improvements: Align orientation events and workshops with practical tasks and ensure better preparation for sessions like the popcorn session. In-person events can enhance engagement and connection.

Further Data Collection: Collect more comprehensive data on performance metrics, workload, and additional factors influencing salaries to refine analysis.

By implementing these recommendations, the financial institution can enhance employee performance, satisfaction, and overall organizational efficiency.

1. Exploratory Data Analysis

We've obtained two HR-related files from the financial institution. The first file, '2016-2024 People Data', contains three tabs: Headcount History, Action History, and Education. The Headcount History consists of 128,572 rows and 10 fields. The Action History has 1,544 rows and 11 fields. Lastly, the Education tab comprises 3,189 rows and 8 fields. Overall, the file contains general HR data, decision history such as promotions, hiring, and retirements, and education information.

The second file, 'Intern Surveys 2018-2021,' includes '18-21 Intern Program Feedback,' '18-21 Midsummer Survey to Mgrs,' and '20-21 Midsummer Survey to Interns.' We've noticed a mix of multiple-choice and open-ended questions. The questions mostly gathered feelings or opinions from interns and their immediate managers post-internship.

2. Our assumption and interests

After reviewing the first file, we've formulated three assumptions as follows:

1) Remote work has a negative impact on performance ratings.

2) Employees with more certificates receive higher annual salaries.

3) Employees with advanced degrees perform better than those with lower-level degrees.

Additionally, we are interested in identifying significant keywords in the extensive text data from the second file. To this end, we decided to identify the most frequently appearing words in each field. Furthermore, we aimed to understand the overall sentiment of managers and interns based on their responses. In brief, we seek to answer the following questions:

4) What are the most common words among managers and interns?

5) What type of sentiment predominates among managers and interns?

3. Out methodologies for data analysis

Regarding the first file, we plan to leverage regression models to identify relationships between relevant variables. For instance, we can use the fields 'location' and 'performance rating' to understand the impact of remote work on performance ratings. Moreover, given the abundance of text data in the second file, we decided to employ NLP methods such as Bag of Words (Term Frequency) and Sentiment Analysis.

4. Regression Analyses on HR data

a. Remote work & performance ratings

First, we used the data from the 'headcount_history' sheet of the '2016-2024 People Data' to find out the relationship between remote work and performance ratings. We decided to use Ordinal Logistic Regression because the dependent variable 'Performance Rating' is ordinal. It allows us to model the probability of observing each possible rating as a function of the mode of work. Additionally, before conducting the ordinal logistic regression, we completed data preprocessing, such as excluding missing data and adjusting the names of locations. Furthermore, we established office work as a baseline for the regression by creating dummy variables. In the 'remote work' field, office work, which includes numerous locations, is coded as 0, and remote work as 1.

As we can see in Appendix 1, we found a statistically significant relationship between working remotely and the odds of receiving a lower performance rating compared to working from the office. If one works remotely, they are 18.3% less likely to receive a higher performance rating. However, we must also consider other variables that might confound or mediate this relationship. Thus, we incorporated 'Annual Salary' data into the model to check whether the relationship changes with the addition of a new variable.

In this case, an employee who works remotely is 9% less likely to receive a higher performance rating. Higher salaries indicate that managers are positioned at higher levels. Thus, the higher their positions, the less their performance ratings are affected by remote work.

In conclusion, we found that remote work can negatively impact performance ratings, but the effect c an also be partially offset by a higher salary or position.

b. Certifications & annual salaries

We attempted to use 'Headcount_history' and 'Education_History'. When merging the two files, we found many unmatched unique identifiers (employee IDs). After excluding the unmatched ones, we decided to focus only on the matched rows. Also, for simplicity, we focused on the data from the year 2023 in the 'Headcount_history'. We wanted to analyze the relationship between the number of certifications (an integer count variable) and annual salaries (considered as a continuous variable) and decided to use a simple linear regression model.

When employees have no certifications, the average annual salary is $110,127, as seen in Appendix 2. Furthermore, although the number of certifications has a statistically significant relationship with salary, it does not account for most of the variability in salaries ($R^2 = 0.044$). Nonetheless, for each additional certification, the annual salary is expected to increase by approximately $4,372 on average, controlling for other factors. The positive coefficient for the number of certifications supports our hypothesis that certifications have a positive effect on annual salary.

Now, to gain a fuller understanding of what affects annual salary, we considered including other relevant variables in the model, such as performance ratings. The second model improved in explanatory power from 4.4% to 9.1%. The expected starting salary with no certifications is $64,358. Moreover, for each additional certification, the annual salary is expected to increase by approximately $4,293, holding the performance rating constant. Conversely, for each one-point increase in the performance rating, the annual salary is expected to increase by approximately $12,767, holding the number of certifications constant. The performance rating appears to have a larger effect on salary than the number of certifications, as indicated by the larger coefficient.

In conclusion, the results support the notion that employees with higher performance ratings and more certifications tend to have higher annual salaries. However, other factors not included in the model also influence annual salary, considering the moderate R-squared value ($R^2 = 0.091$).

c. Education & performance ratings

For a better understanding of our next analysis, we created a dummy variable called 'Degree Level' based on the expected length of years for higher education completion. For example, we set GED and HS as 0 years, but for a bachelor's degree, we assign 4 years for completion. In this way, we quantified various degrees as degree levels: 2 years (A.A., A.A.S., A.S.), 4 years (B.A., B.S.), 5 years (L.L.M.), 6 years (M.A., M.L., M.S., M.B.A., M.P.A., M.P.P.), 7 years (J.D., M.L.A.), 8 years (Ed.D, M.Ed), and 10 years (Ph.D.).

Like in analysis 'b,' we focused on the data from the year 2023 for simplicity. Also, there were some

missing data in the degree field of the 'Education_History,' so we excluded the rows containing missing data.

Regarding our regression model for this case, we are going to use Multinomial Logistic Regression because we suspect issues such as rating leniency or a central tendency bias in the performance rating data. These issues might cause the ratings not to truly reflect an ordinal scale where each increment is of equal weight or importance. Instead, ratings might cluster around certain values or not accurately represent a true hierarchy of performance.

Multinomial Logistic Regression does not assume any order in the outcome categories, so it treats the performance ratings as nominal. This means that each performance rating level will be modeled independently as its own category in relation to a reference category, and the model will estimate the probability of each performance rating level given the predictors.

Based on the result in Appendix 3, for the most part, degree levels do not have a statistically significant association with performance ratings, with the exception of the cases mentioned above. This suggests that the degree level may not be a strong predictor of performance ratings overall, or the effects may vary significantly by degree level.

5. Data Analysis on Surveys

We've tried to figure out what kind of topics are most prominent in interns' or managers' answers to open-ended questions. Although we initially intended to use Topic Modeling for this objective, we found out that it would be hard to gain meaningful results due to the limited answers of around 100. Instead, we've used **'Bag-of-Words'** method to count the frequency of words in each column. In addition, we've implemented **'Sentiment Analysis'** to understand the overall feeling of respondents.

For a better understanding, we've replaced each question with key word. We then found the most five frequent words and overall status of sentiment in the answers.

a. Intern Survey 2018-2021 program feedback

There are twelve questions in the first survey. After excluding the last question due to lack of sufficient answers, we utilized the remaining eleven questions. Based on our sentiment analysis, we've found that the **orientation, events & workshops, and BIE week** were not that strongly positive experience. We've tracked the reasons of this less positive sentiment by using most frequent words in answers of interns.

For the orientation, some of them felt it unnecessary and too long. Moreover, one thought that it could have been more coordinated with supervisors. For events & workshops, we found some room for improvement for the popcorn session. Some interns felt stressed during the session because they hadn't prepared for their questions in advance. Also, it could have been arranged with smaller groups for better engagement. Lastly, as an opinion of minority, we could find that the duration of the BIE week and the presentation time could have been longer to get more meaningful results.

However, experiences with **mentors** were mostly positive. The majority of interns thought that mentors helped them reach out to other colleagues at the financial institution and provided useful resource and advice. Some of interns thought that in-person meetings-not virtual ones- and the frequency of meetings should have increased, though.

Question	Top 5 Frequent Words	Sentiment (Intensity)
Q2. Orientation	Orientation 14, think 9, information 9, long 9, session 8	Positive: 28 (0.60) Neutral: 64 (0.00) Negative: 2 (-0.25)
Q3. Mentors	Mentor 73, bank 17, help 17, meeting 16, great 16	Positive: 57 (0.74) Neutral: 35 (0.00) Negative: 2 (-0.50)
Q8. Events and Workshops	Session 7, popcorn 5, workshop 4, bank 4, good 4	Positive: 10 (0.81) Neutral: 84 (0.00) Negative: 0
Q9. BIE Week	Week 13, bie 12, think 10, experience 9, intern 8	Positive: 14 (0.64) Neutral: 78 (0.00) Negative: 2 (-0.31)

b. Intern Surveys to Inter 2018-2021

Many interns admitted they encountered some challenges. However, we - as data analysts – were hard to find the reasons only by using our textual analysis methodology. So, we've gone over neutral or negative answers manually and found two reasons. First, some interns experienced technical problems like access issues (photoshop, firewall) and new software. Also, they even felt isolated in the virtual environment and wanted to find more people to make their networking.

Additionally, we found out that virtual work had a slightly negative impact on interns' performance. They felt less engaged, connected, and productive while working from home. Although working from home is a double-edged sword like an intern said, there seems to be supplementary measures for better performance.

Question	Top 5 Frequent Words	Sentiment (Intensity)
Q1. Challenges	Work 45, challenge 36, project 27, time 20, manager 18	Positive: 44 (0.61) Neutral: 15 (0.01) Negative: 3 (-0.44)
Q2. Virtual work impacting on performance	Work 37, think 19, people 17, time 15, person 15	Positive: 30 (0.57) Neutral: 26 (0.01) Negative: 6 (-0.43)

c. Intern Surveys to Mgrs 2018-2021

More than 80% of managers felt a positive impression from their interns. The intensity score is 0.82, which means strongly positive, shows that experiences with interns were quite meaningful. A lot of managers mentioned that interns fit their teams well and they also had relevant skills.

However, like experiences of interns, managers thought that virtual work had a negative impacting on the performance of interns. Many of them suggested that working at office can provide more opportunities for engagement and networking for interns.

Question	Top 5 Frequent Words	Sentiment (Intensity)
Q1. Your Intern	Work 80, team 40, well 35, intern 32, skill 28	Positive: 88 (0.82) Negative: 17 (0) Negative: 0
Q2. Virtual work impacting on performance	Work 18, intern 16, bank 15, think 12, well 10	Positive: 22 (0.60) Negative: 79 (0.00) Negative: 4 (-0.29)

6. Additional Data Collection

To better understand performance ratings within the financial institution, you would want to collect a variety of data that captures both the individual characteristics of employees and their work environment. Performance is multifaceted and can be influenced by many factors. For example, we can consider objective performance metrics such as project deliverables, publication record, and policy contributions. In addition, workload and work patterns can affect the performance ratings. In this case, we need to investigate the number of projects handled, report deadlines or work schedules.

For more accurate estimation on annual salary, we can collect additional data such as gender, ethnic diversity, market adjustment, or involvement in high-impact projects & policy making. Especially, as in previous research, the salary gap by gender is still effective in numerous companies. So, it would be meaningful to check whether a similar gap in salary by gender is prevalent at the financial institution.

Finally, the survey data is close to a small-sized one with the number of around one hundred, which is less effective for regression analysis. If we gain more data from a lot of employees, we will be able to come up with a more accurate analysis.

7. Recommendation

Considering the correlation between certifications and annual salary, HR could explore ways to encourage continuous professional development. Investment in certification programs that are relevant to employee roles may not only boost morale but also contribute to their skillset, leading to higher productivity. It is recommended to communicate the value of such programs clearly to employees and possibly integrate them with a well-defined career progression path.

As we saw a negative impact of remote work, the HR department needs to reconsider the continuation of remote work. As there might be pros and cons in this work mode, we will be able to conduct A/B testing to understand potential impacts from remote work. For example, we can split participants into three groups such as 100% remote work, 50% remote work(hybrid), and 100% office work, and then track their tasks (the number of meetings, emails, and so on) during a specific period of time. By doing so, we could find some meaningful insights regarding the optimization of collaboration and performance within the context of the financial institution.

In addition, we found some room for improvement about orientation, events, and BIE week. In overall, events need to be aligned more with practical tasks and frontline employees. In some events like popcorn, interns need to prepare for the events in advance for a better engagement and takeaway. Moreover, events should be organized in-person mode to enhance a sense of connection among interns.

To encapsulate, although insights derived from data offer essential guidance, it is crucial to consider them alongside qualitative feedback and a comprehensive grasp of the organization's dynamics. The financial institution can maintain and enhance a culture of high performance by implementing focused strategies that uphold its operational objectives as well as its employees' welfare.

Appendix

1. Ordinal Logistic Regression (OLR)

[the result of OLR regarding Performance Ratings & Remote work]

Model Summary - Performance Rating

Model	Deviance	AIC	BIC	df	X^2	p
H_0	229465.052	229473.052	229511.784	474116		
H_1	229393.868	229403.868	229452.282	474115	71.184	< .001

Coefficients

	Estimate	Standard Error	z	p
(Intercept) ∗ 1	−6.595	0.079	−83.360	< .001
(Intercept) ∗ 2	−4.396	0.027	−164.899	< .001
(Intercept) ∗ 3	−0.220	0.006	−36.670	< .001
(Intercept) ∗ 4	2.511	0.011	226.687	< .001
remote work1	−0.202	0.024	−8.544	< .001

Note. Performance Rating levels: 1:1, 1:2, 1:3, 1:4, 1:5. Linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3]), logitlink(P[Y<=4]).

[The result of 2nd OLR when adding 'Annual Salary']

Model Summary - Performance Rating

Model	Deviance	AIC	BIC	df	X^2	p
H_0	181517.418	181525.418	181563.213	375132		
H_1	177413.353	177425.353	177482.046	375130	4104.065	< .001

Coefficients

	Estimate	Standard Error	z	p
(Intercept) ∗ 1	−5.352	0.084	−63.903	< .001
(Intercept) ∗ 2	−3.338	0.035	−96.299	< .001
(Intercept) ∗ 3	0.914	0.020	46.831	< .001
(Intercept) ∗ 4	3.736	0.024	156.664	< .001
remote work1	−0.100	0.025	−3.999	< .001
Annual Salary	-1.131×10^{-5}	1.813×10^{-7}	−62.395	< .001

Note. Performance Rating levels: 1:1, 1:2, 1:3, 1:4, 1:5. Linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3]), logitlink(P[Y<=4]).

2. Linear Regression

[Linear Regression – Annual Salary & Number of certifications]

Model Summary - Annual Salary ▼

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	38202.800
H_1	0.209	0.044	0.044	37357.215

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	$9.712\times10^{+11}$	1	$9.712\times10^{+11}$	695.888	< .001
	Residual	$2.118\times10^{+13}$	15177	$1.396\times10^{+9}$		
	Total	$2.215\times10^{+13}$	15178			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	114177.953	310.080		368.221	< .001
H_1	(Intercept)	110126.198	339.899		323.997	< .001
	number of certifications	4371.736	165.724	0.209	26.380	< .001

[Linear Regression – Including the covariate 'Performance rating']

Model Summary - Annual Salary

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	38202.800
H_1	0.302	0.091	0.091	36419.227

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	$2.023\times10^{+12}$	2	$1.011\times10^{+12}$	762.520	< .001
	Residual	$2.013\times10^{+13}$	15176	$1.326\times10^{+9}$		
	Total	$2.215\times10^{+13}$	15178			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	114177.953	310.080		368.221	< .001
H_1	(Intercept)	64358.438	1658.854		38.797	< .001
	number of certifications	4292.530	161.587	0.206	26.565	< .001
	Performance Rating	12766.693	453.403	0.218	28.157	< .001

3. Multinomial Logistic Regression [Education degree & Performance ratings]

Model Summary - Performance Rating

Model	Deviance	AIC	BIC	df	X^2	p
H_0	19745.714	19753.714	19782.597	40408		
H_1	19534.125	19582.125	19755.419	40388	211.590	< .001

Coefficients

	Estimate	Standard Error	z	p
(Intercept) ∗ 1	−17.535	1124.456	−0.016	0.988
(Intercept) ∗ 2	−17.535	1124.456	−0.016	0.988
(Intercept) ∗ 3	2.904	0.296	9.796	< .001
(Intercept) ∗ 4	2.079	0.306	6.791	< .001
Degree level2 ∗ 1	0.065	1209.069	5.386×10^{-5}	1.000
Degree level2 ∗ 2	16.436	1124.456	0.015	0.988
Degree level2 ∗ 3	−0.653	0.321	−2.032	0.042
Degree level2 ∗ 4	−0.234	0.331	−0.705	0.481
Degree level4 ∗ 1	13.431	1124.456	0.012	0.990
Degree level4 ∗ 2	15.040	1124.456	0.013	0.989
Degree level4 ∗ 3	−1.295	0.299	−4.327	< .001
Degree level4 ∗ 4	−0.458	0.309	−1.482	0.138
Degree level6 ∗ 1	−0.311	1356.292	-2.292×10^{-4}	1.000
Degree level6 ∗ 2	16.436	1124.456	0.015	0.988
Degree level6 ∗ 3	−1.597	0.351	−4.551	< .001
Degree level6 ∗ 4	−0.443	0.356	−1.242	0.214
Degree level8 ∗ 1	17.535	10765.584	0.002	0.999
Degree level8 ∗ 2	17.535	10765.584	0.002	0.999
Degree level8 ∗ 3	19.138	7570.778	0.003	0.998
Degree level8 ∗ 4	−2.079	10706.699	-1.942×10^{-4}	1.000
Degree level10 ∗ 1	16.410	3917.751	0.004	0.997
Degree level10 ∗ 2	16.410	3917.751	0.004	0.997
Degree level10 ∗ 3	15.711	1617.149	0.010	0.992
Degree level10 ∗ 4	17.229	1617.149	0.011	0.991

Note. Performance Rating levels: 1:1, 1:2, 1:3, 1:4, 1:5. '5' is the reference level.

4. Bag of Words (Term Frequency)

Install Natural Language ToolKit (NLTK) module (and some other modules)

```
#the module 'sys' allows istalling module from inside Jupyter
import sys

!{sys.executable} -m pip install numpy
import numpy as np

!{sys.executable} -m pip install pandas
import pandas as pd

#Natrual Language ToolKit (NLTK)
!{sys.executable} -m pip install nltk
import nltk

!{sys.executable} -m pip install sklearn
from sklearn import metrics
#from sklearn.model_selection import GridSearchCV
from sklearn.feature_extraction.text import  CountVectorizer #bag-of-words vectorizer
from sklearn.decomposition import LatentDirichletAllocation #package for LDA
```

```
# Plotting tools

from pprint import pprint
!{sys.executable} -m pip install pyLDAvis #visualizing LDA
import pyLDAvis
import pyLDAvis.lda_model

import matplotlib.pyplot as plt
%matplotlib inline

#define text normalization function
%run ./Text_Normalization_Function.ipynb #defining text normalization function

#ignore warnings about future changes in functions as they take too much space
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

Read the CSV file

```
[ ]  # Now use pandas to read the csv file
     program_feedback = pd.read_csv("/content/Intern Surveys 2018-2021 program feedback.csv")
```

```
[ ]  # Read open-ended questions
     business_area_goals_and_work = program_feedback["Business Area Goals and Work"]
     orientation = program_feedback["Orientation"]
     mentors = program_feedback["Mentors"]
     BIE_capstone_project = program_feedback["BIE Capstone Project"]
     presentations = program_feedback["Presentations"]
     most_valuable_experience = program_feedback["Most valuable experience"]
     Recommendation_for_improvement = program_feedback["Recommendation for improvement"]
     events_and_workshops = program_feedback["Events and workshops"]
     BIE_week = program_feedback["BIE Week"]
     favorite_experience = program_feedback["Favorite experience"]
     one_thing_for_improvement = program_feedback["One thing for improvement"]
     onboarding = program_feedback["On-Boarding"]
```

Open-ened Question 1: Business Area Goals and Work

```
# Change 'business_area_goals_and_work' into a list of strings
STR_one_thing_for_improvement = []
for item in one_thing_for_improvement:
    if isinstance(item, float):
        item = str(item)
    STR_one_thing_for_improvement.append(item)
```

+ Code + Text

```
# Eliminate missing values
STR_one_thing_for_improvement = [item for item in STR_one_thing_for_improvement if item != 'nan']
STR_one_thing_for_improvement
```

```
['I wish there was more time for BIE week, as stated above. ',
 'Less down time. I had a short period of my summer where I felt like I had nothing valuable to come in and
 'More relevant popcorn sessions -- the skype session and resume workshops ',
```

Define two functions what will display the results of fitting a topic model.

```
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print("Topic %d:" % (topic_idx))
        print(" ".join([feature_names[i]
                        for i in topic.argsort()[:-no_top_words - 1:-1]]))

def get_topic_words(vectorizer, lda_model, n_words):
    keywords = np.array(vectorizer.get_feature_names_out())
    topic_words = []
    for topic_weights in lda_model.components_:
        top_word_locs = (-topic_weights).argsort()[:n_words]
        topic_words.append(keywords.take(top_word_locs).tolist())
    return topic_words
```

Text Normalization & Vectorization

By normalizing text, we can elmininate stopwords, extend contractions, do stemming or lemmatization.

```
# Normalize text
NORM_Q11 = normalize_corpus(STR_one_thing_for_improvement)
```

Have a look at the Bag-of-Words representation of our corpus

```
df = pd.DataFrame(data = bow_Q11.todense(), columns = bow_vectorizer.get_feature_names_out())
```

+ Code + Text

```
# Assuming 'df' is your DataFrame
column_sums = df.sum()
sorted_sums = column_sums.sort_values(ascending=False)
sorted_sums
```

```
intern      16
week        10
think        9
team         8
work         7
            ..
maybe        1
may          1
many         1
actual       1
workshop     1
Length: 187, dtype: int64
```

5. Sentiment Analysis

Sentiment Analysis

Intall required toolkits.

```
#packages needed

#ignore warnings about future changes in functions as they take too much space
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.filterwarnings("ignore", category=DeprecationWarning)

#the module 'sys' allows istalling module from inside Jupyter
import sys

!{sys.executable} -m pip install numpy
import numpy as np

!{sys.executable} -m pip install pandas
import pandas as pd

#Natrual Language ToolKit (NLTK)
!{sys.executable} -m pip install nltk
import nltk

#text normalization function
%run ./Text_Normalization_Function.ipynb

#ignore warnings about future changes in functions as they take too much space
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

Read the CSV file

```
[2]  # Now use pandas to read the csv file
     program_feedback = pd.read_csv("/content/Intern Surveys 2018-2021 program feedback.csv")
     intern_survey = pd.read_csv("/content/Intern Surveys to Inter 2018-2021.csv")
     manager_survey = pd.read_csv("/content/Intern Surveys to Mgrs 2018-2021.csv")
```

```
[3]  # Read open-ended questions in the program feedback survey
     business_area_goals_and_work = program_feedback["Business Area Goals and Work"]
     orientation = program_feedback["Orientation"]
     mentors = program_feedback["Mentors"]
     BIE_capstone_project = program_feedback["BIE Capstone Project"]
     presentations = program_feedback["Presentations"]
     most_valuable_experience = program_feedback["Most valuable experience"]
     Recommendation_for_improvement = program_feedback["Recommendation for improvement"]
     events_and_workshops = program_feedback["Events and workshops"]
     BIE_week = program_feedback["BIE Week"]
     favorite_experience = program_feedback["Favorite experience"]
     one_thing_for_improvement = program_feedback["One thing for improvement"]
     onboarding = program_feedback["On-Boarding"]

     # Read open-ended questions in the intern survey
     challenges = intern_survey["Challenges"]
     virtual_work_impacting_on_performance_intern = intern_survey["Virtual work impacting on performance"]

     # Read open-ended questions in the manager survey
     your_intern = manager_survey["Your intern"]
     virtual_work_impacting_on_performance_manager = manager_survey["Virtual work impacting on performance"]
```

Normalize the text of answers and name the NORM_Q*

```
[63] # Change 'business_area_goals_and_work' into a list of strings
     STR_virtual_work_impacting_on_performance_manager = []
     for item in virtual_work_impacting_on_performance_manager:
         if isinstance(item, float):
             item = str(item)
         STR_virtual_work_impacting_on_performance_manager.append(item)
```

```
[64] NORM_Q2 = normalize_corpus(STR_virtual_work_impacting_on_performance_manager)
```

Lexicon-Based Sentiment Analysis (Unsupervised Machine Learning)

We will use the VADER lexicon available through the NLTK module.

+ Code + Text

```
[6]  nltk.download('vader_lexicon')

     from nltk.sentiment.vader import SentimentIntensityAnalyzer
     analyzer = SentimentIntensityAnalyzer()
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
```

Let's score all answers in each open-ended question.

```
def analyze_sentiment_vader_lexicon(answer, threshold = 0.1, verbose = False):
    scores = analyzer.polarity_scores(answer)

    # Adjust the binary sentiment assignment to include 'neutral' category
    if scores['compound'] >= threshold:
        binary_sentiment = 'positive'
    elif scores['compound'] <= -threshold:
        binary_sentiment = 'negative'
    else:
        binary_sentiment = 'neutral'

    if verbose:
        print('VADER Polarity (Binary):', binary_sentiment)
        print('VADER Score:', round(scores['compound'], 2))
    return binary_sentiment,scores['compound']
```

```
[65] VADER_polarity_test = [analyze_sentiment_vader_lexicon(answer, threshold=0.1) for answer in NORM_Q2]
     VADER_polarity_test_df = pd.DataFrame(VADER_polarity_test, columns = ['VADER Polarity','VADER Score'])
     VADER_polarity_test_df.head()
```

	VADER Polarity	VADER Score
0	neutral	0.0000
1	neutral	0.0000
2	neutral	0.0000

```
# Count the occurrences of each polarity
polarity_counts = VADER_polarity_test_df['VADER Polarity'].value_counts()

# Calculate the average score for each polarity
average_scores = VADER_polarity_test_df.groupby('VADER Polarity')['VADER Score'].mean()

# You can then print or store the results as needed
print(polarity_counts)
print(average_scores)
```

```
VADER Polarity
neutral     79
positive    22
negative     4
Name: count, dtype: int64
VADER Polarity
negative   -0.288675
neutral     0.000000
positive    0.597055
Name: VADER Score, dtype: float64
```

[The result of Sentiment Analysis]

Question	Top 5 Frequent Words	Sentiment (Intensity)
Q1. Business Area Goals and Work	Work 59, Project 18, Team 16, Like 16, learn 13	Positive: 47 (0.75) Neutral: 44 (0.00) Negative: 3 (-0.24)
Q2. Orientation	Orientation 14, think 9, information 9, long 9, session 8	Positive: 28 (0.60) Neutral: 64 (0.00) Negative: 2 (-0.25)

Q3. Mentors	Mentor 73, bank 17, help 17, meeting 16, great 16	Positive: 57 (0.74) Neutral: 35 (0.00) Negative: 2 (-0.50)
Q4. BIE Capstone Project	Project 48, work 44, think 25, week 23, group 22	Positive: 38 (0.70) Neutral: 54 (-0.01) Negative: 2 (-0.29)
Q5. Presentations	Presentation 33, bank 17, enjoy 17, like 12, intern 11	Positive: 44 (0.74) Neutral: 50 (0.00) Negative: 0
Q6. Most Valuable Experience	Work 33, experience 28, bank 23, valuable 19, learn 18	Positive: 41 (0.68) Neutral: 53 (0.00) Negative: 0
Q7. Recommendation for Improvement	Intern 38, work 21, think 18, bie 17, time 16	Positive: 34 (0.66) Neutral: 57 (0.01) Negative: 3 (-0.36)
Q8. Events and Workshops	Session 7, popcorn 5, workshop 4, bank 4, good 4	Positive: 10 (0.81) Neutral: 84 (0.00) Negative: 0
Q9. BIE Week	Week 13, bie 12, think 10, experience 9, intern 8	Positive: 14 (0.64) Neutral: 78 (0.00) Negative: 2 (-0.31)
Q10. Favorite experience	People 14, bank 13, intern 11, work 11, employee 8	Positive: 23 (0.68) Neutral: 71 (0) Negative: 0
Q11. One thing for improvement	Intern 16, week 10, think 9, team 8, work 7	Positive: 14 (0.63) Neutral: 79 (-0.00) Negative: 1 (-0.30)