# 1. Introduction

### Background and Problem Statement

Foundations Project Bank (FPB) is a relatively new and rapidly growing challenger online bank which caters to retail banking customers.

I am a Data Analyst and I have been tasked with determining whether or not FPB has been compliant with fair salary regulations as part of a review that was called for in a recent shareholder meeting.

The analysis I performed used HR data to determine whether there was a gender pay (gap) bias at FPB.

I have chosen to focus on the gender pay gap as it is very topical in banking at the moment. An article by the Financial Times, "UK banks among companies with the largest pay gaps" from April 5 2019 highlighted this issue.

Given this is a concern for regulators and future customers any perception by the public that FPB is not complying with fair salary regulations may affect shareholder value by putting off potential customers, creating issues with our regulators and perhaps putting of female employees from joining and staying with FPB.

### Stakeholder management

Below is a list of FPBs stakeholders to be considered as part of this review:

- Shareholders
- Regulators
- Press and Media
- Customers
- Board of Directors
- Management
- FPB employees

### Data Protection including Ethical and Legal concerns

As part of my analysis I used HR Data which was made available to me in its entirety. While this Data allowed me to conduct my analysis as I do have some concerns to raise

- The HR data is freely available to me and anyone else in the organisation to analyse.
- The data contains personal detail including Full names, dates of birth Social security numbers, passwords and addresses.
- The fact that employee passwords are available in the HR data may also mean that customer data and sensitive or classified information and data is at risk of being fraudulently accessed and viewed.
- Not all of this data is required by all employees and there are no safe guards to ensure personal data is not accessed by unauthorised employees or without justified cause.
- It is not clear whether the employees of FPB are aware that their data is freely available.
- It is not clear that there are any controls around who can access personal data and when they can do so or even the format in which data can be viewed.
- All of the above poses not only ethical concerns but legal concerns as is would appear FPB do not have an adequate Data protection policy or is complying with laws including Data Protection and or GDPR.

## 2. Problem Solution and recommendations

**Compliance with fair pay regulations**

- Management should use the result of this analysis as an input to a proactive response to any adverse findings
- FPB must get ahead of the curve by acknowledging any possible bias in gender pay and any failure to comply with Fair pay regulations
- A pay review must be put in place with a view to introducing a transparent pay scale and introducing a time scale which details when the bank will close and gender pay gap

**Ethics and legal concerns**

- Immediately introduce a review into data protection at FPB. A data protection policy must be introduced at FPB and all data be subject to classification to protect customers and employees.
- HR and other data must be protected and must be compliant with Data Protection and GDPR regulations
- Data must only be made available must be for a clearly defined purpose. For example there is no reason for a Data Analyst to be provided with the passwords of all employees

## 3. Implementation Part 1 (please review alongside the Jupyter Notebook download)

**Working with others**

I have worked with one very helpful employee who allowed me to include their personal data to demonstrate the lack of controls over personal information.

As shown in section 3.2 of the code I have provided the personal information that was in the HR Data set I was given access to for my analysis. This includes Social security Numbers, Passwords and Phone Numbers

**Tasks completed in producing the report**

**3.1** Importing the Python libraries to be used in the statistical analysis of the HR Data

**3.2** Loading and checking the HR data. This is an initial high level check of the data provided by HR including the types of data. This ensures that the data in a "clean" state and is ready for analysis.

**3.3** Redacting the data. The data provided by HR contains private data including Social security Numbers, Passwords and Phone Numbers. This detail is not required so I have redacted the data set. As outlined in the "Working with others" section I have kept one row of data visible to demonstrate the personal data that I was able to access and the legal and ethical issues this could pose for FPB.

**3.4** Summarising the HR Data

It is worth noting that although I use various tools and statistical methods to analyse the HR data our stakeholders may not require much evidence and rigorous testing as I have performed to form a negative perception of FPB.

The perception alone of a gender bias is enough to lasting damage the FPB brand. As such the inferences we can arrive at in summarising the HR data are of great importance.

Per the points below and in my analysis file and code the average salary for men is 77 481 while the average salary for women is 55 837. The highest salary in the company is paid to a man - 118 826 for men vs 89 450 for the highest paid woman. The lowest salary paid to a woman is 25 583 vs 30 014 for men. The higher standard deviation figure in the male salaries does suggest that the salaries are less grouped around the average salary of the company

- Using .describe() to summarise the HR data by Gender and Salary. This allows an initial high level view of the salaries of Male and Female employees.
- Using seaborne to visualise the differences in salary between men and women. I used the seaborn distribution plot function (sns.distplot) because it allows us to visualise the salary ranges of men and women as well as their densities of alongside each other.
- Visualising the Salary data spread using Boxplot (seaborn). I added this second visualisation as it complimented the variance and standard deviation details included in my analysis. The boxplot shows that Male employees while having a wider interquartile range also have higher salaries on average and a higher maximum salary amount

## 4. Implementation Part 2 : Defining and Testing my hypotheses (Please review alongside the Jupyter Notebook download)

Defining my hypothesis: $\mathbf{H_0}$ : There is no Employment bias based on Gender (pay) at FPB (e.g there is no difference the salaries of Male and Female employees) and $\mathbf{H_A}$ : There is an Employment bias based on Gender (pay) at FPB. Significance level : 95% ($p < 0.05$)

**4.1** Creating a subset of the "Salary" column by Gender and separating my data into Male_Salary and Female_salary categories.

**4.2** Testing for normality in our Male salaries and Female salaries. I have done this to help me decide the most appropriate test to use in my analysis.

The normality tests and the histograms show that both the male and the female salaries are not normally distributed. This is because the p-values from both the Male and Female salary normality tests are both close to zero.

Given the data is not normally distributed for both male and female salaries I have decided to use the "mann_whitney" test as makes fewer assumptions about data being normally distributed and equal variances in the data

**4.3** Using function mann_whitney test to see if there are differences between the salaries of men and women. The p-value is 1.2082264089084797e-05 so the observed differences are very unlikely to be due to chance. I reject the null hypothesis which is that there is no Employment bias based on Gender (pay) at FPB (e.g there is no difference the salaries of Male and Female employees

**4.4** An alternative test. Using function stats.ttest to see if there are differences between the salaries of men and women. Although the stats.ttest does make assumptions which are not true in our data set e.g. that variances are equal Female (19487) Male (23788) I have included this test in my analysis because the stakeholders may not give the same attention to the test selected that I have done. In this case reputational and brand damage can still be done to FPB even if an inappropriate statistical test is used.

In this test The p-value was : 2.7650476599728173e-06 The observed differences are very unlikely to be due to chance, we reject the null hypothesis

## 5. Conclusion

I have conducted a range of tests to understand our HR data with a focus on the salaries of Male and Female employees.

I have used various Python libraries including pandas and seaborn and statistical methods to analyse the HR Data set. I have checked the data and redacted the personal data I did not need for my analysis.

Based on the fact that male and female salaries are not normally distributed at FPB I used the mann_whitney test. In this test the p-value is 1.2082264089084797e-05 so the observed differences are very unlikely to be due to chance.

On this basis I reject the null hypothesis and conclude that there is an employee bias at FPB which is creating a situation where Male employees are paid more than Female employees. In this respect the fears raised by our shareholders appear to hold true.

As I mentioned in the report, I am a data analyst and I have followed the techniques I have learnt. One of the concerns I raised was that this HR Data is freely available to anyone in the organisation and while I have chosen the mann_whitney test because it has relaxed assumptions about normality of data and variance anyone else analysing the same data set may not do the same. As such I have included the stats.test. This test does arrive at the same conclusion. The observed differences are very unlikely to be due to chance.

The findings are not positive for any of our stakeholders and need to be urgently addressed in order to avoid a negative impact on the business.

In addition while useful for my analysis the data set I was provided also poses legal and ethical risks for the business as the personal details and passwords on all our employees are freely available to anyone. For example the fact that everyone with access to the HR data can view the passwords of any of our employees may mean that even the personal details of our customers can easily be exposed which in itself threatens our ability to be allowed to conduct business by our regulators.

While the issues raised in this analysis are of great concern the actions and recommendations are clear decisions and actions need to be taken with regards to the existing

I. Gender pay gap and the data and
II. Legal and ethical issues around HR Data

## 6. Evaluation

The results and analysis I have done provides the shareholders with the detail they have requested. It would appear that there is a Gender pay gap at FPB and in addition there are legal and ethical issues to address on how our HR data is stored and accessed.

This analysis will have helped the business if the business acts to remediate the risks I have highlighted. The impact/risk of not acting on these issues is large and can possibly affect our business and brand going forward.

I suggest that the same analysis I have performed on Gender pay be replicated for Age, Ethnicity and Marital status bias and Goals and actions be set to address any issues that are found to exist.

Finally I am happy to perform this analysis again in the future to review any progress the business has made to address the issues I have raised.