HDSC Summer'22 Capstone Project Presentation: Gender Statistics – Earnings of Male and Female Employees

A Project by Team Data Warehouse

INTRODUCTION

The Australian labor market is highly gender-segregated by industry and occupation, a pattern that has persisted over the past two decades. The gender wage gap is an indicator of women's earnings compared to men and it is derived by the difference between the average annual earnings for women and that of men. According to the International Labor Organization (ILO), on average, women globally are paid 20% less than men. The ILO further attributes discrimination based on gender as the largest contributory factor to the pay gap. Data Science techniques can be used in an organization to identify wage disparities between male and female employees, hence serving as an inference for companies to examine their hiring practices, policies and resolve gender pay gaps.

PROBLEM STATEMENT

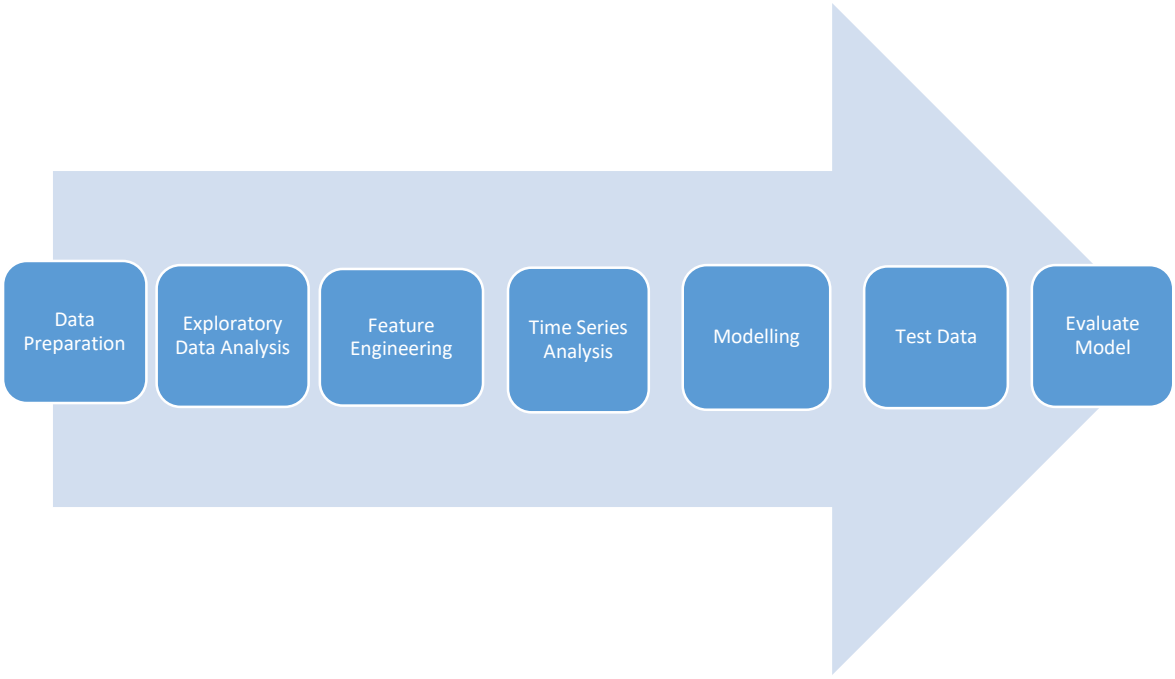
The Workplace Gender Equality Agency (WGEA) estimates that Australian women earn, on average, 15.3% less than men. That is $1,387 for women per week compared with $1,638 for men. To bridge this gap, the Australian government has committed huge funds to sponsor more women across different fields in recent years, but despite this sponsorship, female average earnings are still significantly less than their male counterparts. The Data Warehouse team of the Hamoye Summer 2022 Internship seeks to explore and gain significant insights from this disparity.

AIMS AND OBJECTIVES

The aim of this project is to

1. Explore and analyze the gender pay disparity of male and female employees in different job roles between 2004 to 2017 in Australia
2. Build a machine learning model that predicts the future wage gap and its impact on the female workforce.

FLOW PROCESS



DATA COLLECTION

The data used in this analysis was collected and downloaded from Kaggle at https://www.kaggle.com/datasets/mpwolke/cusersmarildownloadsearningcsv and sent to the project repository.

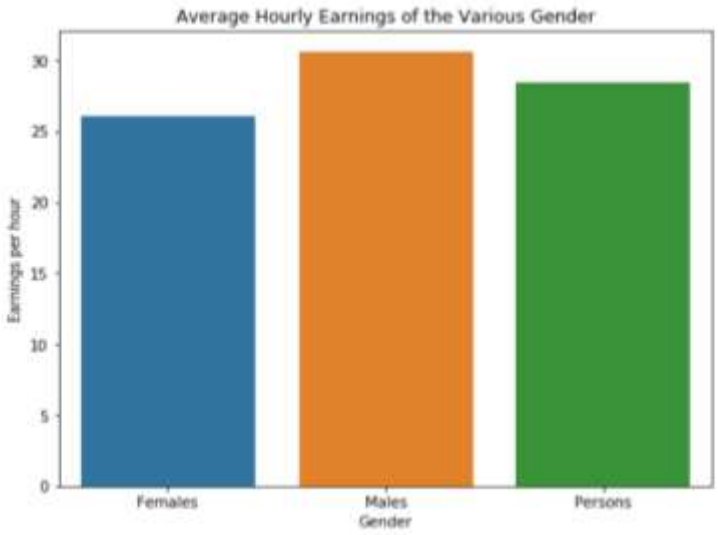
DATA PRE-PROCESSING

The data set used in the analysis is a CSV file containing the average hourly earnings of females, males and persons, across different job categories from 2004 to 2017; with 14 rows and 28 columns.

The data was cleaned and manipulated by renaming its columns with a proper naming convention and was found to have no missing or duplicate values. The data frame was subdivided into three according to gender for further analysis.

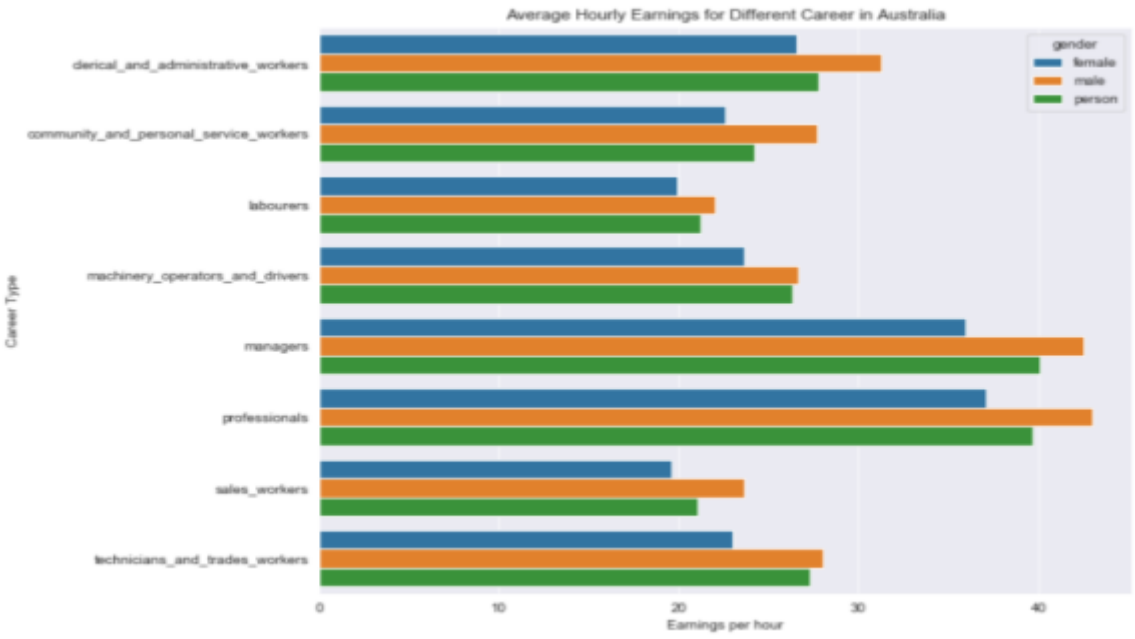
EXPLORATORY DATA ANALYSIS

Exploratory data analysis was carried out to understand the trend of female and male earnings over the years and reveal the job positions with the highest male-to-female earning disparities.

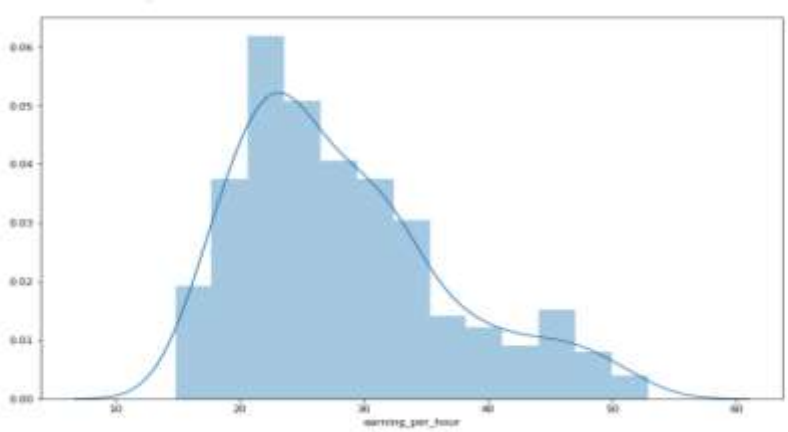


A bar chart of Hourly earnings vs. Gender

The figure above shows the average hourly earnings of the eight cadres over the total period of analysis (2004 - 2017) for the various genders. The male gender receives the highest hourly wage of 30.59 AUD while the female gender receives the least of 26.05 AUD. Generally, hourly earnings increased regardless of gender or the type of job over the years.



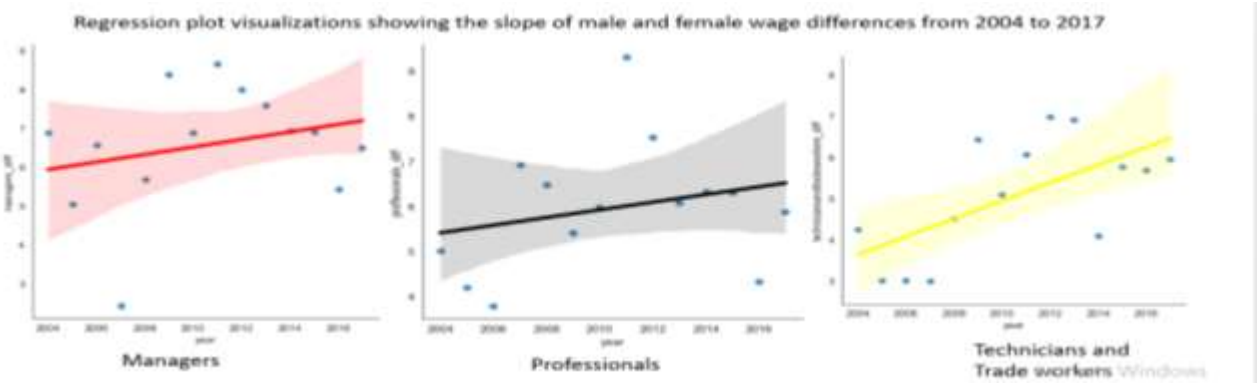
The Plot above shows how hourly wages differ across the various careers and it further reveals that the male gender has the highest hourly pay while the female has the least. Also, the difference between male and the other genders tend to increase for careers with more hourly pay. It is observed that managers and professionals are the most paid job titles.



From the visualization above, it can be seen that overall, irrespective of cadre, most employees in Australia earn an average hourly wage of about 25 AUD; also an hourly minimum wage of 15AUD.

REGRESSION ANALYSIS

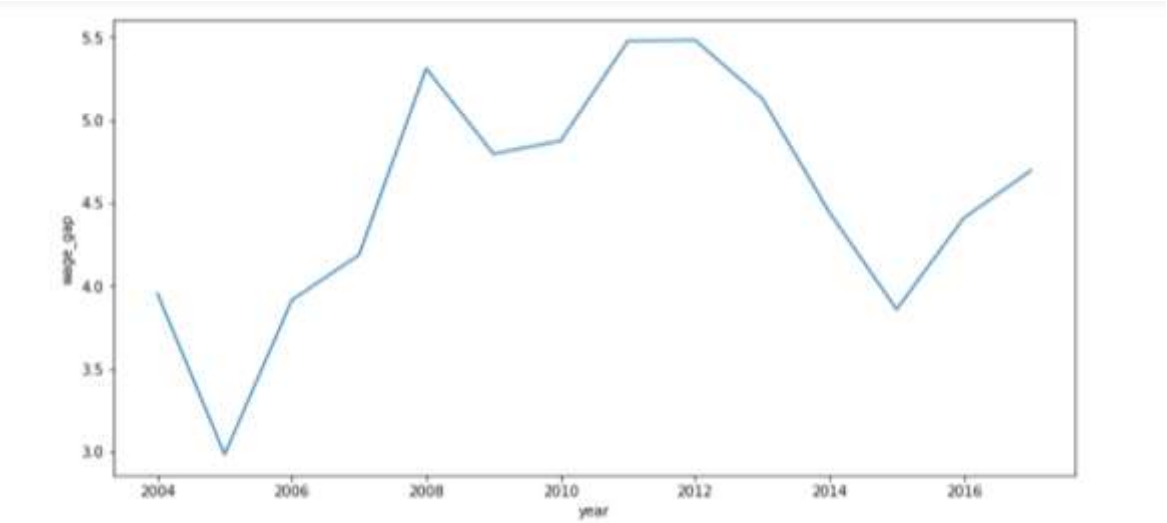
The male and female data were extracted from the data to avoid dropping the 'persons' data. The person's data stems from people who did not specify their gender: including them will be irrelevant to our analysis. Eight data frames containing the years of consideration and the male and female in that cadre were created; also, the difference between the male and female earnings was computed and plotted using the line plot from seaborn. It was observed that some plots have steep slope (the increase in their wage gap per unit increase in the female and/or male earnings over the years); while others were more gentle. Notable were the machinery operators and drivers who have negative slope.



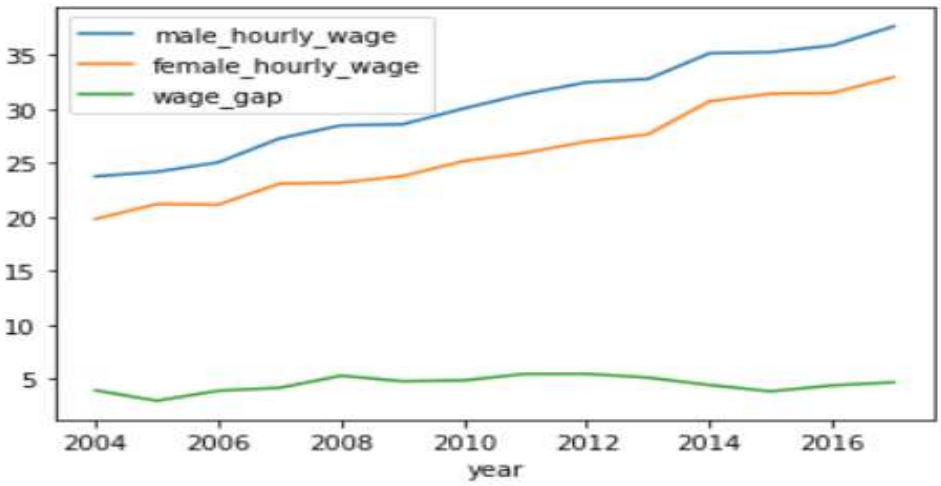
The plots above show the relationship between male and female earnings with respect to job titles. In the relationship, the difference for most jobs is positive meaning that the male earns more that the female counterpart.

FEATURE ENGINEERING

Engineering the features involved recreating the data frame such that the annual average hourly earnings of male and female irrespective of career was made into two columns: “male_hourly_data” and “female_hourly_data”; this collapsed the data frame into one entry per year. The target variable (wage gap) was calculated from the difference between the male and female earnings for each year in the dataset. This was done by querying out the male and female data and then creating an average hourly wage each year for both males and females. Then, the hourly wage gap between males and females was created. The wage gap was then plotted against each year to achieve the diagram below.



A line graph of wage gap versus year.



We realize that the increase in wage gap has been on a steady trend around 5 dollars in hourly wage over the years. We can easily infer that this data doesn't have seasonality. The wage gap is independent of time.

MODEL TRAINING AND EVALUATION

The model used a Time Series analysis and modeling (The Box - Jenkins Method) to predict the future wage gap. The analysis conducted through the depicture of the ACF and PACF plots, respectively, resulted in the selection of the ARIMA model order (0, 0, 0), i.e. q and p. The ARIMA and Prophet algorithms were used to model the data, of which the better model was the former; with Mean Square Error and Mean Absolute Percentage Error of 0.16 and 0.07 respectively.

CONCLUSION AND RECOMMENDATIONS

The dataset had limited features and was modeled on time series. The ARIMA model was able to determine future wage gap, further confirming the inequality in earnings. The evidence from this project revealed that, the average hourly wages for men across all job titles were greater than those for women in Australia. However, in the years 2005 and 2015, there was an exceptional event where women who worked as drivers and machinery operators made more money. This could have been attributed to the strong employment growth during these periods according to the Australian Government Treasury Report and Employment in Australia.

One of the challenges faced while modeling the data was insufficiency of independent features (variables) to reliably forecast our target variable without bias, and probably make recommendations on factors that can be put into consideration to reduce the wage gap disparity.

We recommend that:

- To automate and predict the gap across all industries, data on the main causes of the wage gap should be made public.
- Formal policies and/or strategies that support gender equality should be in place inside organizations.
- Top cadres have lower proportion of women; hence, more women ought to be inspired to enroll in university professional programs.
- To determine whether and where discrepancies may exist, the employer should always do a pay gap analysis for all the roles, and perhaps the reason for the salary disparity and how to close it.
- Consistent efforts should be made to lessen the impact of the major causes of the pay disparity.