Informative Title

Ava Cascario, Sydney Holt, and Kacie Rohn

2024-12-03

```
# Load Packages
library(ggplot2)
library(dplyr)

# Load Data
careerDataRaw <- read.csv(
  file = "career_change_prediction_dataset.csv",
  header = TRUE,
  sep = ","
)
# Clean data by removing rows
careerDataCleaned <- careerDataRaw[-c(13:23)]
# Summarize data for data visualization
careerDataOccupationGender <- careerDataCleaned %>%
  select(
    Current.Occupation, Gender
  ) %>%
  group_by(
    Current.Occupation, Gender
  ) %>%
  summarize(
    Count = n(),
    .groups = 'drop'
  )

# Create bar chart displaying Gender distribution by Current Occupation
ggplot(
  data = careerDataOccupationGender,
  aes(
    x = Current.Occupation,
    y = Count,
    fill = Gender
  )
) +
  geom_bar(
    stat = "identity",
```

```
    position = position_dodge(width = 0.8)
  ) +
  theme_minimal() +
  scale_fill_manual(
    values = c(
      "Male" = "lightblue",
      "Female" = "pink"
    ),
    name = "Gender"
  ) +
  labs(
    title = "Gender Distribution by Current Occupation",
    x = "Current Occupation",
    y = "Count"
  ) +
  theme(
    axis.text.x = element_text(size = 7, angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, size = 16)
  )
# Summarize Job Satisfaction by Field of Study and Current Occupation
jobSatisfactionSummary <- careerDataCleaned %>%
  group_by(Field.of.Study, Current.Occupation) %>%
  summarize(
    count = n(),
    min = min(Job.Satisfaction),
    Q1 = quantile(Job.Satisfaction, 0.25),
    median = median(Job.Satisfaction),
    Q3 = quantile(Job.Satisfaction, 0.75),
    max = max(Job.Satisfaction),
    medianAbsoluteDeviation = mad(Job.Satisfaction),
    sampleArithmeticMean = mean(Job.Satisfaction),
    SampleArithmeticSD = sd(Job.Satisfaction),
  )

print(jobSatisfactionSummary)
```

Research Topic: Field of Study vs Occupation

Our focus in conducting our study is to look at the association between field of study and actual job occupation. That is, to know if what an individual studied would have an affect on the job they get post-graduation. We will conduct our research by looking at different research questions and creating visualizations to represent the data that correlates with each question. We will then explore our outcomes and summarize our findings and how they connect back to our topic.

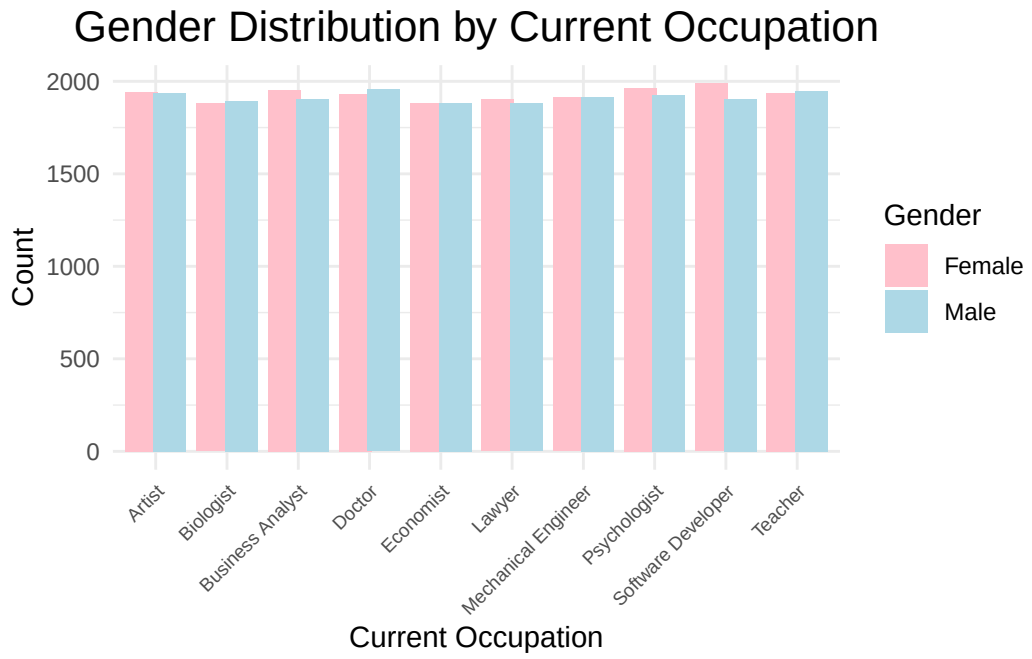
Research Questions

The first question we will explore is, how does your field of study correlate with your job occupation? We want to know what kind of association is found between the two factors, and if what an individual studied has an affect on the job they pursue. We would also like to explore, what the most commonly chosen occupations are based on the field of study? For example, if an individual was a software engineering major, what is the most commonly chosen job after graduation. Finally, we would like to explore the question, is there an association between gender and field of occupation? Not only will our research focus on field of study and occupation, but we would also like to analyze whether gender plays a role in this as well.

Provenance Of Our Data

We are utilizing a data set that we found on Kaggle. Kaggle is a website focused towards data scientists with a goal in helping others learn about data. The data was collected by Jahnavi Paliwal, a data science masters student at the University of San Francisco. The data has a usability score of 10.0 and is annually updated by Jahnvai, with the latest update being one month ago. This means the data is collected from a reliable source, and updated consistently to remain relevant. The data set is designed to help you predict whether individuals are likely to change their occupation based on their academic background, job experience, and other demographic factors. This can help you answer questions based on numerous aspects of the job industry including human resources, income, industry analysis, job markets, and job availability. This data set contains over 30,000 records each with 22 attributes. This data set constitutes a case as a single individual.

FAIR and CARE Principles



```
# A tibble: 100 x 11
# Groups:   Field.of.Study [10]
   Field.of.Study Current.Occupation  count   min    Q1 median    Q3   max
   <chr>          <chr>               <int> <int> <dbl>  <dbl> <dbl> <int>
 1 Arts           Artist                352     1     3    6       8    10
 2 Arts           Biologist             397     1     3    5       8    10
 3 Arts           Business Analyst      366     1     3    5.5     8    10
 4 Arts           Doctor                399     1     3    5       8    10
 5 Arts           Economist             365     1     3    6       8    10
 6 Arts           Lawyer                360     1     3    5       8    10
 7 Arts           Mechanical Engineer   381     1     3    5       8    10
 8 Arts           Psychologist          390     1     3    6       8    10
 9 Arts           Software Developer    383     1     3    6       8    10
10 Arts           Teacher               370     1     3    6       8    10
# i 90 more rows
# i 3 more variables: medianAbsoluteDeviation <dbl>,
#   sampleArithmeticMean <dbl>, SampleArithmeticSD <dbl>
```