Project1

Bhavya Patel

2024-09-29

About Dataset

I have found a data set on "AI-Powered Job Market Insights" from Kaggle.
This data set focuses on the role of AI and automation across various industries. This data set includes 500 unique job listings, each characterized by industry, company size, AI adoption level, automation risk, required skills, and job growth projections.
More information about the dataset can be found here: https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights

Problem

Question: Is there a correlation between AI adoption level, Automation Risk, and Salary? Do jobs in industries with higher AI adoption levels offer better pay?

To solve this problem, I would do a correlation analysis to check any relationship between AI adoption level, automation risk, and salary. A linear regression model will determine the impact of AI adoption level and automation risk on salary.

Let's look at some of the data:

```
unique(df$Industry)
```

```
##  [1] "Entertainment"      "Technology"         "Retail"
##  [4] "Education"          "Finance"            "Transportation"
##  [7] "Telecommunications" "Manufacturing"      "Healthcare"
## [10] "Energy"
```

```
unique(df$Job_Title)
```

```
##  [1] "Cybersecurity Analyst" "Marketing Specialist"  "AI Researcher"
##  [4] "Sales Manager"         "UX Designer"           "HR Manager"
##  [7] "Product Manager"       "Software Engineer"     "Data Scientist"
## [10] "Operations Manager"
```

```
summary(df$Salary_USD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31970   78512   91998   91222  103971  155210
```

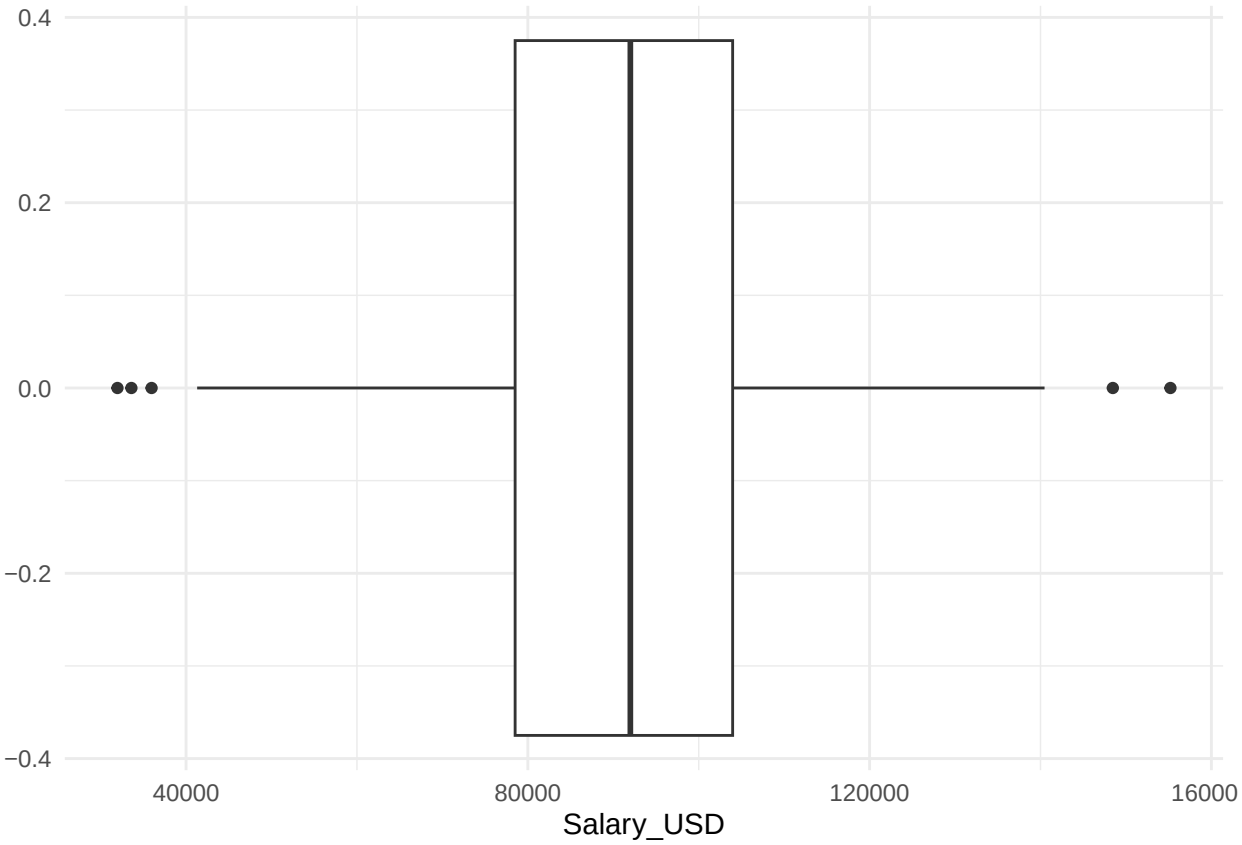
Data Wrangling and Cleaning

Let's start by cleaning the data, removing any duplicate records.

```
df <- df %>% distinct()
```

Let's make box plot for salary and remove all very low paying jobs from the data set.

```
ggplot(df, aes(x = Salary_USD)) +
  geom_boxplot() + theme_minimal()
```




```
df <- df %>% filter(Salary_USD >= 40000)
```

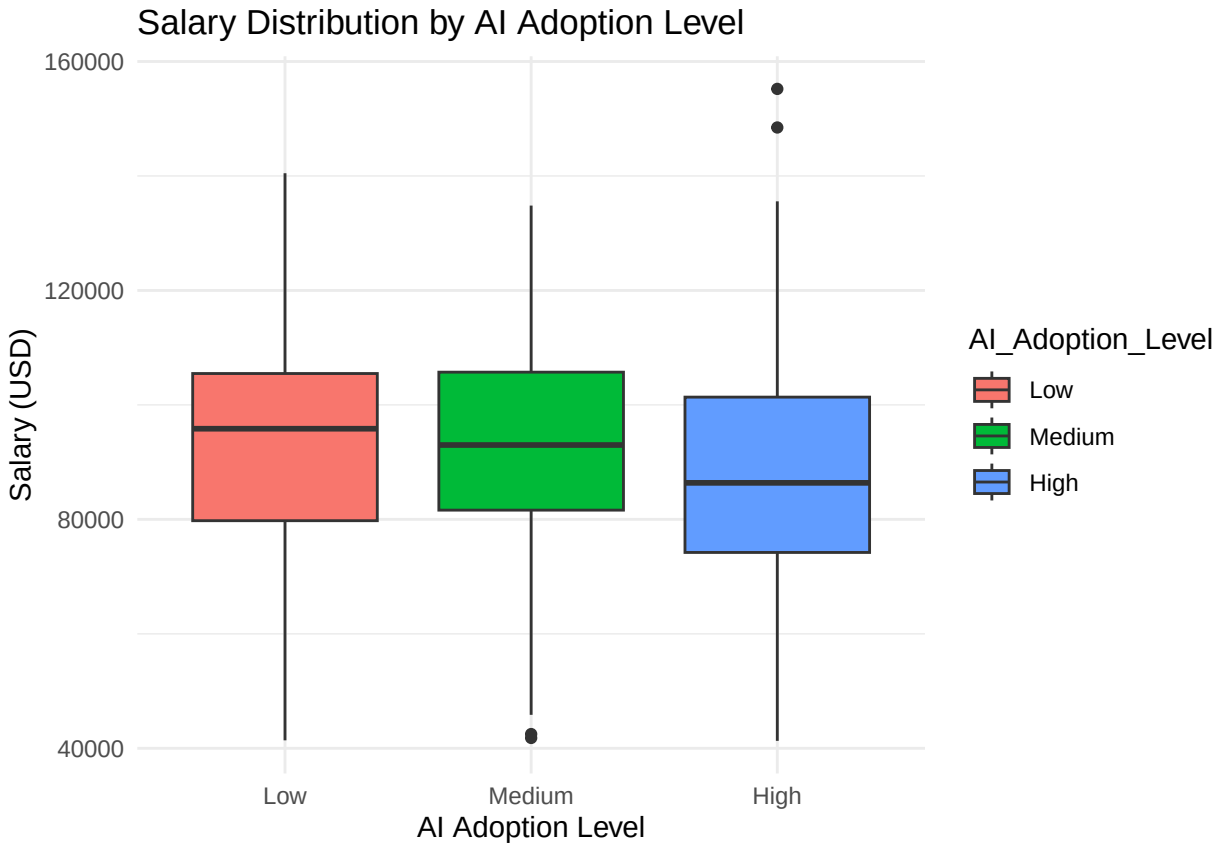
Converting categorical data into factors.

```
df$AI_Adoption_Level <- factor(df$AI_Adoption_Level, levels = c("Low", "Medium", "High"))
df$Automation_Risk <- factor(df$Automation_Risk, levels = c("Low", "Medium", "High"))
```

Data Exploration

Salary Distribution by AI Adoption Level

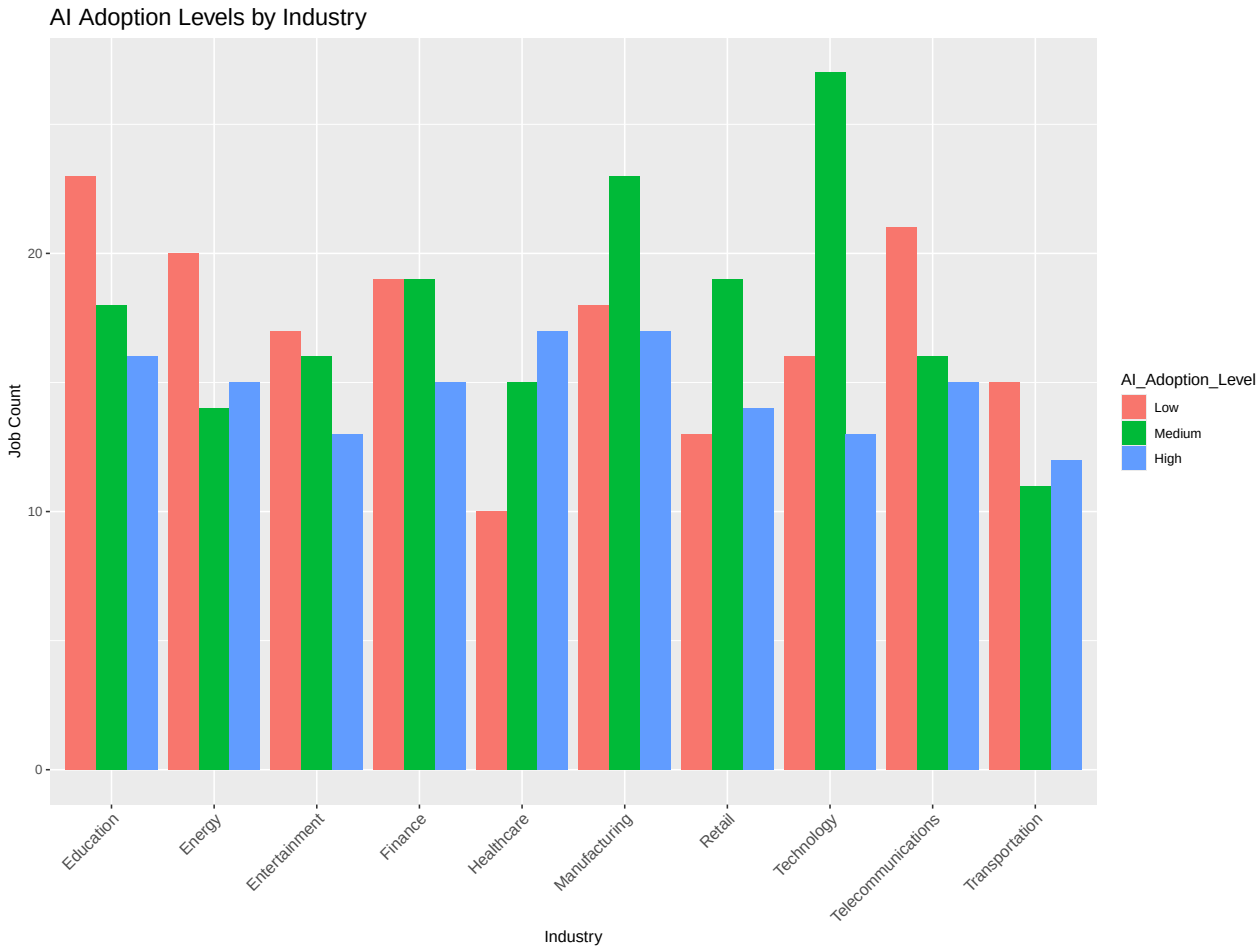
```
ggplot(df, aes(x = AI_Adoption_Level, y = Salary_USD, fill = AI_Adoption_Level)) +
  geom_boxplot() +
  labs(title = "Salary Distribution by AI Adoption Level",
       x = "AI Adoption Level", y = "Salary (USD)") +
  theme_minimal()
```



The Average Salary got slightly decreased by in increase in adoption level. But there are some jobs in high AI adoption levels that pays more than the any other jobs.

AI Adoption across industries

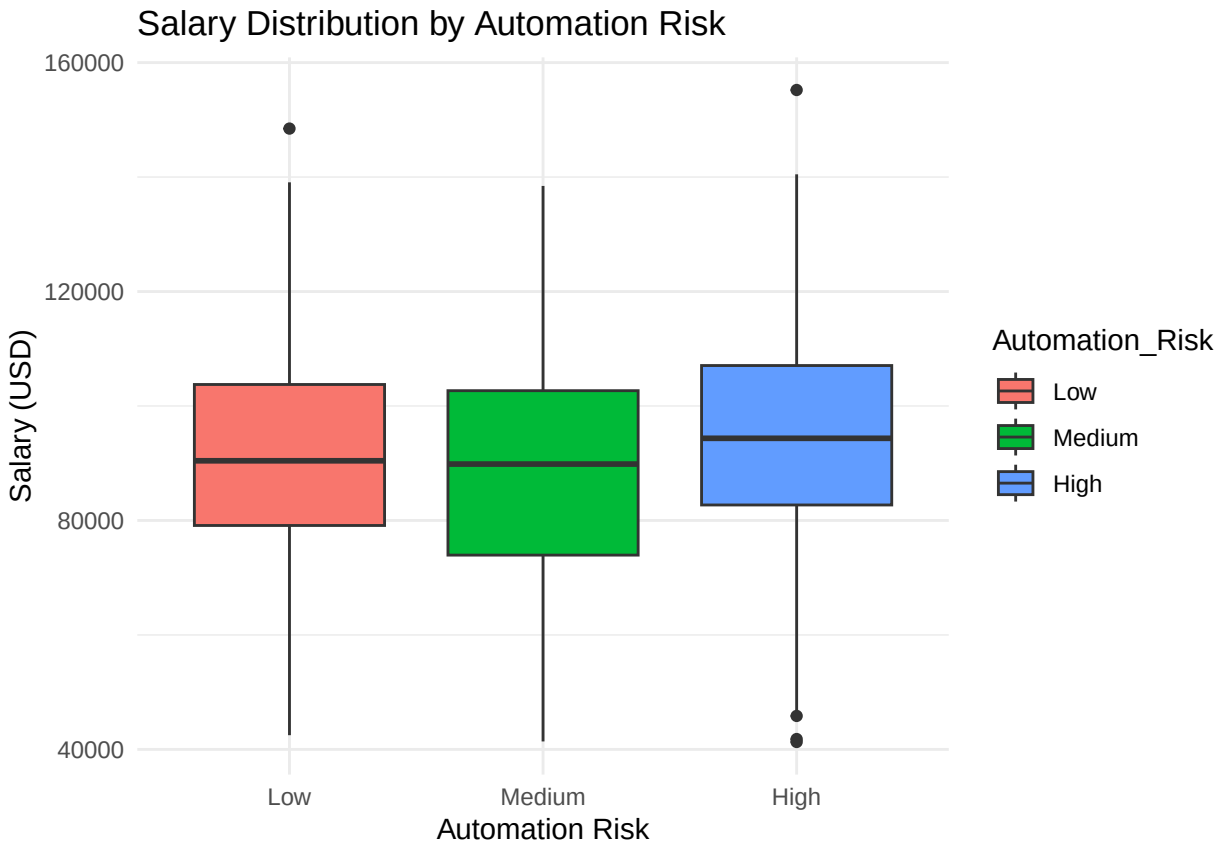
```
ggplot(df, aes(x = Industry, fill = AI_Adoption_Level)) +
  geom_bar(position = "dodge") +
  labs(title = "AI Adoption Levels by Industry",
       x = "Industry", y = "Job Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 11), plot.title = element_text(size=16)
```



From the graph, Education, Healthcare, and Manufacturing have the most jobs in high AI adoption levels. Overall Technology has most jobs in all AI adoption levels.

Salary Distribution by Automation Risk

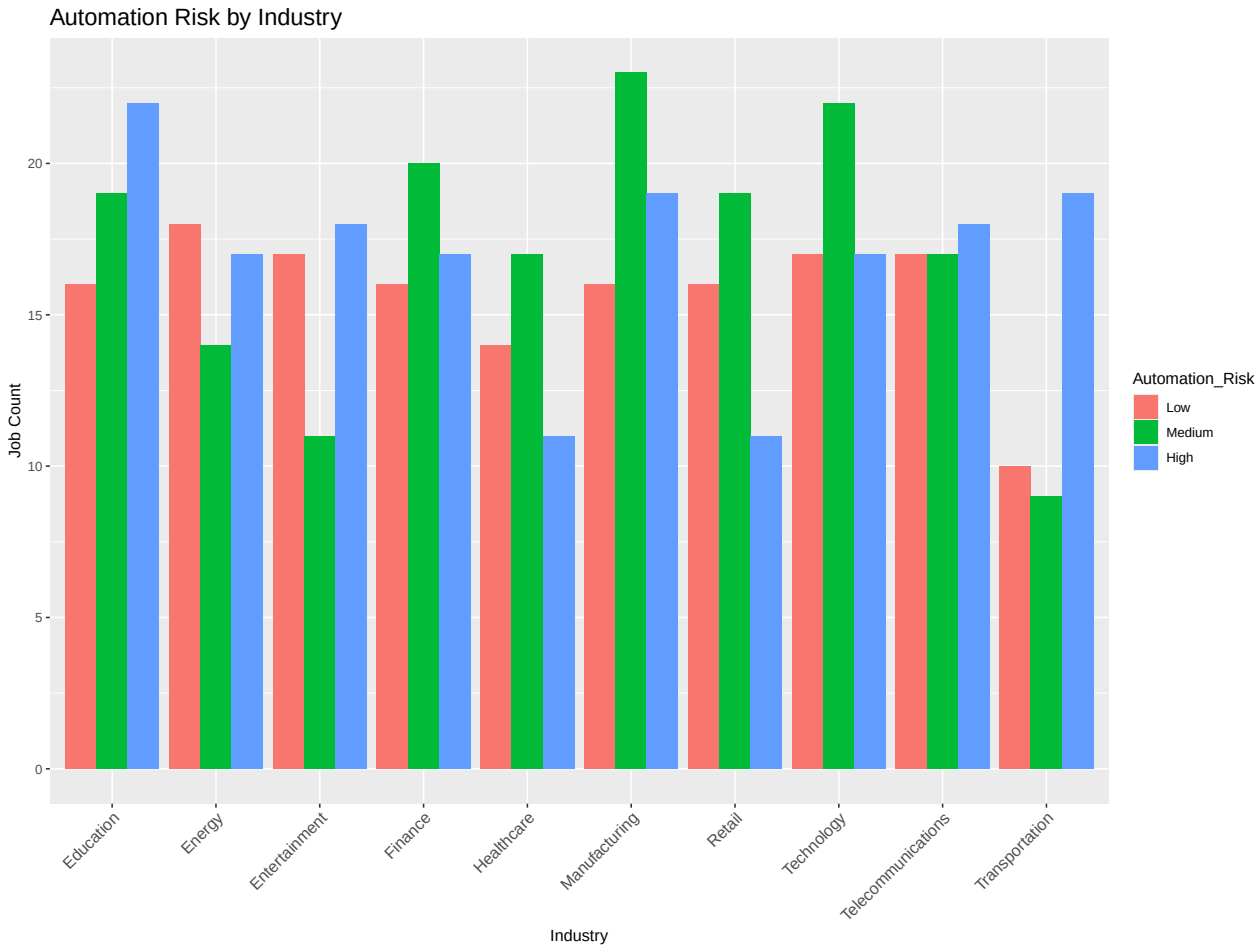
```
ggplot(df, aes(x = Automation_Risk, y = Salary_USD, fill = Automation_Risk)) +
  geom_boxplot() +
  labs(title = "Salary Distribution by Automation Risk",
       x = "Automation Risk", y = "Salary (USD)") +
  theme_minimal()
```



This graph shows that Salary distribution based on Automation Risk is almost similar in all levels. Some jobs in high risk have very low salaries.

Automation Risk across industries

```
ggplot(df, aes(x = Industry, fill = Automation_Risk)) +
  geom_bar(position = "dodge") +
  labs(title = "Automation Risk by Industry",
       x = "Industry", y = "Job Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 11), plot.title = element_text(size=16
```



Almost all Industry have high risk of automation except for Healthcare and retail. Education, Manufacturing, and Transportation have most jobs that are at highest risk automation.

Top 5 Skills

```
df_skills <- df %>%
  separate_rows(Required_Skills, sep = ",") %>%
  group_by(Required_Skills) %>%
  summarise(Skill_Count = n()) %>%
  arrange(desc(Skill_Count))
top_skills <- df_skills[1:5, ]
knitr::kable(top_skills, format = "markdown", col.names = c("Skills", "Job Count"))
```

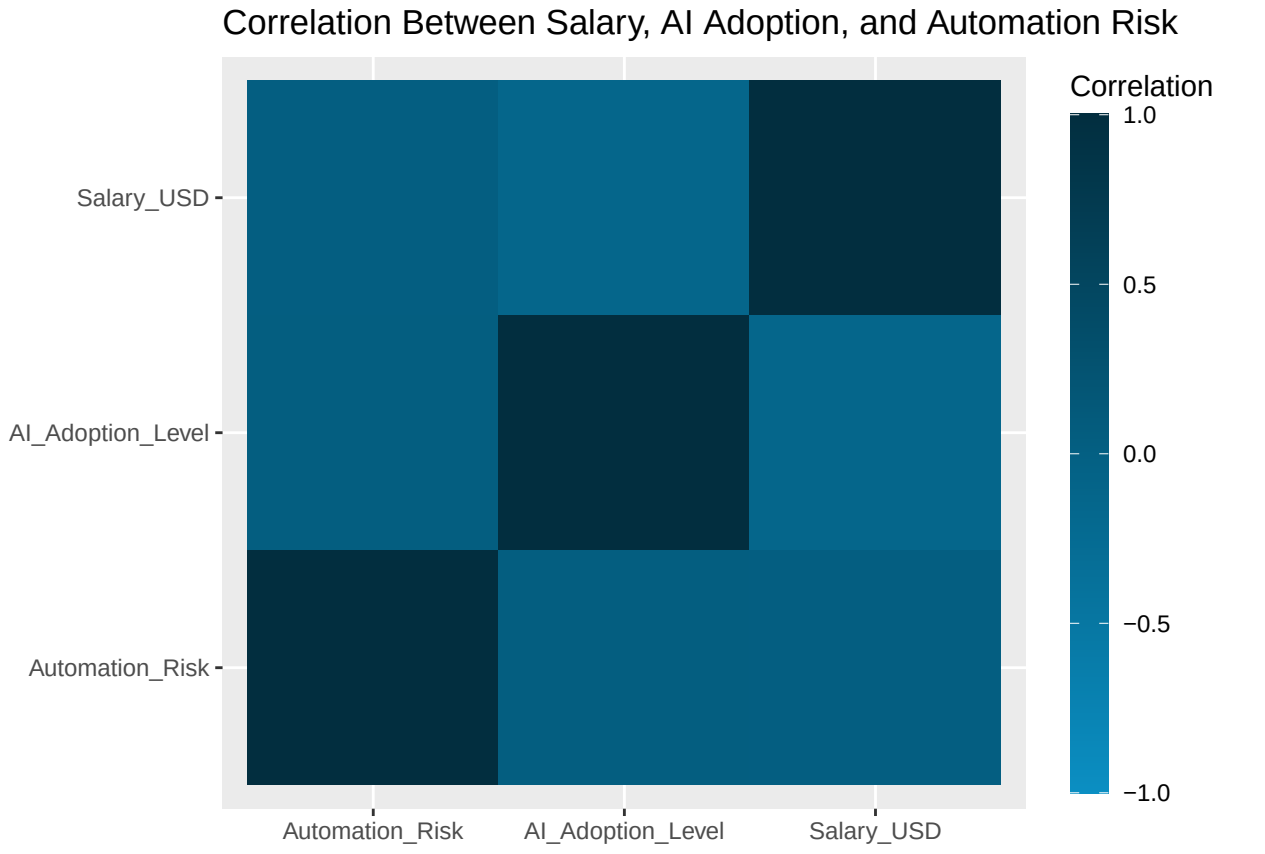
Skills	Job Count
Project Management	60
Python	59
Cybersecurity	58
Machine Learning	52
Sales	49

Data Analysis

Correlation Analysis

Correlation analysis is a statistical method used to evaluate the strength and direction of relationships between two or more variables. In this study, we analyzed the correlation between three key variables: Automation Risk, AI Adoption Level, and Salary in USD. The correlation coefficients, which range from -1 to 1, provide insights into how these variables interact with one another.
Since AI adoption and automation risk are categorical, we'll use Cramér's V for categorical-categorical correlations and point-biserial correlation for categorical-numerical correlations.

```
df_numeric <- df %>%
  mutate(AI_Adoption_Level = as.numeric(as.factor(AI_Adoption_Level)),
         Automation_Risk = as.numeric(as.factor(Automation_Risk)))
cor_matrix <- cor(df_numeric[, c("Automation_Risk", "AI_Adoption_Level",
"Salary_USD")], use = "complete.obs")
cor_data <- melt(cor_matrix)
ggplot(cor_data, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "#0B8FC3", high = "#022E3F", mid = "#046083",
midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
  labs(title = "Correlation Between Salary, AI Adoption, and Automation Risk",
       x=NULL, y=NULL) +
  guides (fill = guide_colorbar(barwidth = 1,  barheight = unit(90, "mm"),
      ticks = TRUE, label = TRUE))
```



```r
kable((cor_matrix), format = "markdown", digits = 4)
```

	Automation_Risk	AI_Adoption_Level	Salary_USD
Automation_Risk	1.0000	0.0392	0.0310
AI_Adoption_Level	0.0392	1.0000	-0.1272
Salary_USD	0.0310	-0.1272	1.0000

Interpretation of Graph & Table:

Automation_Risk correlation with Salary_USD: -0.1136, suggesting a weak negative correlation between automation risk and salary. Higher automation risk is associated with a small decrease in salary.

Automation_Risk correlation with AI_Adoption_Level: -0.0607, indicating a very weak and negligible negative relationship. A slight decrease in AI adoption corresponds to a minimal increase in automation risk.

AI_Adoption_Level correlation with Salary_USD: 0.0925, suggesting a weak positive correlation. As AI adoption levels rise, salary tends to increase slightly, though the relationship is very minimal.

The correlation analysis conducted provides a foundational understanding of the relationships between Automation Risk, AI Adoption Level, and Salary. Although the observed correlations are weak.

Next let's do regression analysis, that allows us to explore and quantify the relationship further. It can help determine how much one variable (predictor) affects another variable (outcome) while controlling for additional factors.

Linear Regression

We will analyze how AI Adoption Level and Automation Risk (independent variables) influence Salary in USD (dependent variable).

```
model <- lm(Salary_USD ~ AI_Adoption_Level + Automation_Risk,
        data = df_numeric)
summary(model)
```

```
##
## Call:
## lm(formula = Salary_USD ~ AI_Adoption_Level + Automation_Risk,
##     data = df_numeric)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53215 -12340    770  13266  66160
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        96054.7     3191.8  30.095  < 2e-16 ***
## AI_Adoption_Level  -3226.5     1119.3  -2.883  0.00412 **
## Automation_Risk      891.6     1105.5   0.807  0.42033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19940 on 494 degrees of freedom
## Multiple R-squared:  0.01748,    Adjusted R-squared:  0.01351
## F-statistic: 4.396 on 2 and 494 DF,  p-value: 0.01282
```

Interpretation of the Output:
Residual Standard Error: 19,910 indicates the average distance that the observed salaries fall from the regression line. This value suggests moderate variability in salary predictions.

The coefficient for AI Adoption Level is 2,138 with a p-value of 0.0546. This suggests that for each one-unit increase in the AI Adoption Level, salaries increase by an average of $2,138, holding Automation Risk constant.
The p-value is close to the 0.05 significance level, indicating that while the relationship is not statistically significant at the 5% level, it is borderline significant.
The coefficient for Automation Risk is -2,626 with a p-value of 0.0155, indicating a statistically significant negative relationship. For each one-unit increase in Automation Risk, the salary decreases by an average of $2,626, holding AI Adoption Level constant. This result is significant at the 0.05 level, indicating strong evidence that higher automation risks are associated with lower salaries.
This negative association may suggest that industries or jobs with higher automation risks could potentially lead to lower compensation, possibly due to fears of job displacement or market adjustments in such sectors.

Multiple R-squared: 0.02026 indicates that approximately 2.03% of the variance in Salary is explained by the model. This low value suggests that the model does not capture much of the variation in salary and that other factors may be influencing salaries that are not included in the model.
The p-value of 0.00638 suggests that there is strong evidence against the null hypothesis, which states that all regression coefficients are equal to zero (i.e., none of the predictors have an effect on the dependent variable). Since this p-value is less than 0.05, we can conclude that the model is statistically significant.

Conclusion

The analysis aimed to explore the relationship between AI Adoption Level, Automation Risk, and Salary to address the initial question.
The correlational analysis provided an initial understanding of the relationships among the variables:
Automation Risk and Salary: A weak negative correlation was found, indicating that as automation risk increases, salary tends to decrease slightly.
AI Adoption Level and Salary (r = 0.0925): A weak positive correlation was observed between AI adoption and salary, indicating that industries or sectors with higher levels of AI adoption tend to offer marginally better pay.
Automation Risk and AI Adoption Level (r = -0.0607): The correlation between automation risk and AI adoption is very weak and negative, suggesting little to no meaningful relationship between these two factors.
While the correlations identified are weak, they provide a preliminary indication of the directions in which the relationships move.

The regression analysis deepened our understanding by quantifying the impact of AI Adoption Level and Automation Risk on Salary, while controlling for the influence of each variable.
The overall model is statistically significant with a p-value of 0.00638, meaning that the predictors collectively have a significant impact on salary. However, the Multiple R-squared of 0.02026 indicates that the model explains only about 2.03% of the variance in salary. This suggests that other factors, not included in this model, are likely influencing salary to a much greater extent.

The combined analysis suggests that both AI Adoption Level and Automation Risk have some influence on Salary, but their effects are relatively weak.
AI Adoption Level shows a weak, borderline significant positive relationship with salary, hinting that industries with higher AI adoption may offer slightly better pay, though this effect is not particularly strong.
Automation Risk has a statistically significant negative relationship with salary, suggesting that jobs with higher automation risks may offer lower compensation, likely due to economic pressures or job insecurity related to automation.

Therefore, while there is a correlation between these factors and salary, the practical significance of these relationships is limited.
Jobs in industries with higher AI adoption levels does offer slightly better pay but no significant difference.