Assgn4

2025-04-14

PROBLEM 1:

I started by reading the *Fashion\_Retail\_Sales.csv* dataset into a Pandas DataFrame and converted the date column to a proper datetime format. After removing rows with missing values in “Purchase Amount (USD)” and “Date Purchase,” I aggregated daily sales by summing the purchase amounts. Next, I created a time series plot using Matplotlib to visualize how total purchase amounts varied over the observed dates. The final graph shows daily total sales over time with notable spikes and dips.

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("Fashion\_Retail\_Sales.csv")

df['Date Purchase'] = pd.to\_datetime(df['Date Purchase'], format="%d-%m-%Y", errors='coerce')

df\_clean = df.dropna(subset=['Purchase Amount (USD)', 'Date Purchase'])

daily\_sales = df\_clean.groupby('Date Purchase')['Purchase Amount (USD)'].sum().reset\_index()

plt.figure(figsize=(12,6))

plt.plot(daily\_sales['Date Purchase'], daily\_sales['Purchase Amount (USD)'], marker='o', linestyle='-')

plt.title('Time Series of Total Purchase Amount (USD)')

plt.xlabel('Date')

plt.ylabel('Total Purchase Amount (USD)')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

A graph of blue lines

Description automatically generated

From the time series graph, the sales do not appear to strictly trend upward or downward overall, but there are noticeable fluctuations. Some days show substantial peaks, indicating periods of potentially high-volume sales (e.g., promotional events or holidays), while other days remain relatively low. These spikes could be tied to seasonal promotions or specific marketing campaigns. Missing data (about 650 rows) might slightly impact the clarity of the trend, so addressing data quality issues would be valuable in future analyses.

**Future Decisions and Trends**

1. **Seasonal Promotions**: If the spikes are linked to specific recurring dates or events, scheduling targeted promotional campaigns around those times could boost revenues.
2. **Inventory Planning**: Sharp increases in daily sales suggest the need to ensure adequate stock before anticipated high-demand periods.
3. **Data Improvement**: Improving data collection processes (to reduce missing values) will make trend analysis and forecasting more reliable.

**Conclusion**  
In summary, the time series plot reveals fluctuating daily sales with occasional dramatic spikes. By investigating the causes behind these spikes—be they seasonal or promotional—the company can refine its marketing efforts, adjust inventory, and plan strategically for peak sales periods. Ensuring better data completeness and continuously monitoring the time series will further improve decision-making and help capture emerging trends in customer purchases.

## PROBLEM 2:

## R Markdown

I began by loading the necessary libraries and importing the employee\_turnover.csv data into R. I inspected the dataset’s structure using functions such as head() and str(), and removed any rows with missing values for work experience (stag) and the turnover indicator (event). I then converted key variables into the appropriate formats—setting work experience and turnover events as numeric and other variables like gender and industry as factors—to ensure accurate analysis. Using the survival package, I created a survival object based on work experience and turnover events, and fitted a Kaplan–Meier model stratified by gender to analyze the time to employee turnover. I visualized the resulting survival curves with ggsurvplot, which included a risk table that detailed the number of employees at risk over time. Additionally, I performed a Cox Proportional Hazards regression to evaluate how variables such as gender, age, industry, and personality traits (like extraversion and selfcontrol) potentially influence the risk of an employee leaving. This analysis not only identifies key risk factors affecting employee turnover but also provides insights for informing future decisions—such as implementing targeted interventions and adjusting compensation policies—to improve employee retention.

(!require("survival")) install.packages("survival")

## Loading required package: survival

if (!require("survminer")) install.packages("survminer")

## Loading required package: survminer

## Warning: package 'survminer' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Loading required package: ggpubr

## Warning: package 'ggpubr' was built under R version 4.3.3

##   
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':  
##   
## myeloma

# Load necessary libraries  
library(survival)  
library(survminer)  
  
# Read CSV  
df <- read.csv("employee\_turnover-2.csv")  
  
# Inspect data  
head(df)

## stag event gender age industry profession traffic coach  
## 1 7.030801 1 m 35 Banks HR rabrecNErab no  
## 2 22.965092 1 m 33 Banks HR empjs no  
## 3 15.934292 1 f 35 PowerGeneration HR rabrecNErab no  
## 4 15.934292 1 f 35 PowerGeneration HR rabrecNErab no  
## 5 8.410678 1 m 32 Retail Commercial youjs yes  
## 6 8.969199 1 f 42 manufacture HR empjs yes  
## head\_gender greywage way extraversion independ selfcontrol anxiety novator  
## 1 f white bus 6.2 4.1 5.7 7.1 8.3  
## 2 m white bus 6.2 4.1 5.7 7.1 8.3  
## 3 m white bus 6.2 6.2 2.6 4.8 8.3  
## 4 m white bus 5.4 7.6 4.9 2.5 6.7  
## 5 f white bus 3.0 4.1 8.0 7.1 3.7  
## 6 m white bus 6.2 6.2 4.1 5.6 6.7

str(df)

## 'data.frame': 1129 obs. of 16 variables:  
## $ stag : num 7.03 22.97 15.93 15.93 8.41 ...  
## $ event : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ gender : chr "m" "m" "f" "f" ...  
## $ age : num 35 33 35 35 32 42 42 28 29 30 ...  
## $ industry : chr "Banks" "Banks" "PowerGeneration" "PowerGeneration" ...  
## $ profession : chr "HR" "HR" "HR" "HR" ...  
## $ traffic : chr "rabrecNErab" "empjs" "rabrecNErab" "rabrecNErab" ...  
## $ coach : chr "no" "no" "no" "no" ...  
## $ head\_gender : chr "f" "m" "m" "m" ...  
## $ greywage : chr "white" "white" "white" "white" ...  
## $ way : chr "bus" "bus" "bus" "bus" ...  
## $ extraversion: num 6.2 6.2 6.2 5.4 3 6.2 6.2 3.8 8.6 5.4 ...  
## $ independ : num 4.1 4.1 6.2 7.6 4.1 6.2 6.2 5.5 6.9 5.5 ...  
## $ selfcontrol : num 5.7 5.7 2.6 4.9 8 4.1 4.1 8 2.6 3.3 ...  
## $ anxiety : num 7.1 7.1 4.8 2.5 7.1 5.6 5.6 4 4 7.9 ...  
## $ novator : num 8.3 8.3 8.3 6.7 3.7 6.7 6.7 4.4 7.5 8.3 ...

# Convert variables to appropriate types  
df$stag <- as.numeric(df$stag)   
df$event <- as.numeric(df$event)   
df$gender <- as.factor(df$gender)   
df$industry <- as.factor(df$industry)  
df$profession <- as.factor(df$profession)  
df$coach <- as.factor(df$coach)  
df$head\_gender <- as.factor(df$head\_gender)  
  
# Remove rows with missing stag or event  
df <- df[!is.na(df$stag) & !is.na(df$event), ]  
  
### Kaplan–Meier Survival Analysis  
  
# Create survival object  
surv\_obj <- Surv(time = df$stag, event = df$event)  
  
# Fit Kaplan–Meier curve stratifying by gender  
km\_fit <- survfit(surv\_obj ~ gender, data = df)  
  
# Plot survival curves with risk table  
km\_plot <- ggsurvplot(  
 km\_fit,  
 data = df,  
 risk.table = TRUE,  
 pval = TRUE,  
 conf.int = TRUE,  
 title = "Survival Analysis of Employee Turnover by Gender",  
 xlab = "Work Experience (months)",  
 ylab = "Probability of Remaining Employed",  
 palette = "Dark2"  
)  
  
# Print the Kaplan–Meier plot  
print(km\_plot)

A graph of a graph of a person and person

Description automatically generated with medium confidence

### Cox Proportional Hazards Regression  
  
cox\_model <- coxph(  
 surv\_obj ~ gender + age + industry + profession + greywage +  
 extraversion + independ + selfcontrol + anxiety + novator,  
 data = df  
)  
  
# View summary of the Cox model  
summary(cox\_model)

## Call:  
## coxph(formula = surv\_obj ~ gender + age + industry + profession +   
## greywage + extraversion + independ + selfcontrol + anxiety +   
## novator, data = df)  
##   
## n= 1129, number of events= 571   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## genderm -0.114189 0.892090 0.124201 -0.919 0.357894   
## age 0.024711 1.025019 0.006421 3.848 0.000119 \*\*\*  
## industryAgriculture 0.713553 2.041231 0.537340 1.328 0.184199   
## industryBanks 0.327156 1.387017 0.429983 0.761 0.446742   
## industryBuilding 0.271455 1.311872 0.453583 0.598 0.549527   
## industryConsult 0.284166 1.328653 0.444166 0.640 0.522319   
## industryetc 0.034530 1.035133 0.439028 0.079 0.937311   
## industryIT -0.476444 0.620988 0.448484 -1.062 0.288081   
## industrymanufacture -0.280707 0.755249 0.432403 -0.649 0.516222   
## industryMining -0.016104 0.984025 0.507958 -0.032 0.974708   
## industryPharma -0.160869 0.851404 0.519912 -0.309 0.757005   
## industryPowerGeneration -0.269900 0.763456 0.492820 -0.548 0.583923   
## industryRealEstate -1.216829 0.296168 0.625095 -1.947 0.051579 .   
## industryRetail -0.328194 0.720224 0.425895 -0.771 0.440945   
## industryState -0.033112 0.967431 0.469700 -0.070 0.943799   
## industryTelecom -0.667064 0.513213 0.497909 -1.340 0.180333   
## industrytransport -0.084426 0.919039 0.485839 -0.174 0.862043   
## professionAccounting -0.132903 0.875550 0.514439 -0.258 0.796140   
## professionBusinessDevelopment 0.413183 1.511621 0.398683 1.036 0.300030   
## professionCommercial 0.753575 2.124582 0.404171 1.864 0.062252 .   
## professionConsult 0.497005 1.643791 0.412766 1.204 0.228557   
## professionEngineer 0.977953 2.659009 0.436105 2.242 0.024931 \*   
## professionetc 0.350894 1.420336 0.382620 0.917 0.359100   
## professionHR 0.198334 1.219370 0.314345 0.631 0.528078   
## professionIT 0.079168 1.082386 0.375022 0.211 0.832808   
## professionLaw 0.208796 1.232194 0.558233 0.374 0.708382   
## professionmanage 0.953379 2.594463 0.399558 2.386 0.017029 \*   
## professionMarketing 0.599300 1.820844 0.386089 1.552 0.120607   
## professionPR 0.496186 1.642444 0.563821 0.880 0.378837   
## professionSales 0.371267 1.449571 0.350055 1.061 0.288873   
## professionTeaching 0.375835 1.456207 0.467546 0.804 0.421486   
## greywagewhite -0.560862 0.570717 0.131961 -4.250 2.14e-05 \*\*\*  
## extraversion 0.030680 1.031155 0.035005 0.876 0.380796   
## independ -0.010033 0.990017 0.034682 -0.289 0.772367   
## selfcontrol -0.038595 0.962140 0.036158 -1.067 0.285792   
## anxiety -0.042716 0.958184 0.033509 -1.275 0.202398   
## novator 0.004681 1.004692 0.029797 0.157 0.875181   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## genderm 0.8921 1.1210 0.69934 1.1380  
## age 1.0250 0.9756 1.01220 1.0380  
## industryAgriculture 2.0412 0.4899 0.71205 5.8516  
## industryBanks 1.3870 0.7210 0.59714 3.2217  
## industryBuilding 1.3119 0.7623 0.53926 3.1914  
## industryConsult 1.3287 0.7526 0.55634 3.1731  
## industryetc 1.0351 0.9661 0.43782 2.4474  
## industryIT 0.6210 1.6103 0.25783 1.4957  
## industrymanufacture 0.7552 1.3241 0.32361 1.7626  
## industryMining 0.9840 1.0162 0.36361 2.6631  
## industryPharma 0.8514 1.1745 0.30732 2.3588  
## industryPowerGeneration 0.7635 1.3098 0.29060 2.0057  
## industryRealEstate 0.2962 3.3765 0.08699 1.0084  
## industryRetail 0.7202 1.3885 0.31257 1.6595  
## industryState 0.9674 1.0337 0.38531 2.4290  
## industryTelecom 0.5132 1.9485 0.19341 1.3618  
## industrytransport 0.9190 1.0881 0.35464 2.3817  
## professionAccounting 0.8755 1.1421 0.31944 2.3998  
## professionBusinessDevelopment 1.5116 0.6615 0.69196 3.3022  
## professionCommercial 2.1246 0.4707 0.96215 4.6914  
## professionConsult 1.6438 0.6083 0.73198 3.6914  
## professionEngineer 2.6590 0.3761 1.13112 6.2508  
## professionetc 1.4203 0.7041 0.67097 3.0066  
## professionHR 1.2194 0.8201 0.65851 2.2579  
## professionIT 1.0824 0.9239 0.51899 2.2574  
## professionLaw 1.2322 0.8116 0.41258 3.6800  
## professionmanage 2.5945 0.3854 1.18561 5.6774  
## professionMarketing 1.8208 0.5492 0.85434 3.8807  
## professionPR 1.6424 0.6088 0.54396 4.9592  
## professionSales 1.4496 0.6899 0.72991 2.8788  
## professionTeaching 1.4562 0.6867 0.58243 3.6408  
## greywagewhite 0.5707 1.7522 0.44065 0.7392  
## extraversion 1.0312 0.9698 0.96278 1.1044  
## independ 0.9900 1.0101 0.92496 1.0597  
## selfcontrol 0.9621 1.0393 0.89631 1.0328  
## anxiety 0.9582 1.0436 0.89728 1.0232  
## novator 1.0047 0.9953 0.94770 1.0651  
##   
## Concordance= 0.637 (se = 0.013 )  
## Likelihood ratio test= 121.6 on 37 df, p=6e-11  
## Wald test = 123.2 on 37 df, p=3e-11  
## Score (logrank) test = 128.7 on 37 df, p=5e-12

## Explanation

From the Kaplan–Meier survival curve, I observed that the survival probability (i.e., the probability of remaining employed) generally decreases over time as expected. When comparing male and female employees, there is some difference in their likelihood of leaving the company at various points, but based on the p-value (displayed on the plot), the difference may or may not be statistically significant (in my example, the p-value was around 0.13, which is above the conventional 0.05 threshold).

The risk table under the survival curve helps me see how many employees remain in the analysis at each time interval. This is particularly useful for gauging how the sample size changes over time.

In the Cox Proportional Hazards model, the output provided hazard ratios (HR), confidence intervals, and p-values for each variable included in the model. For example, if greywage had an HR significantly greater than 1, it would suggest that employees with lower compensation (slightly above minimum wage) might be more likely to leave. If gender showed an HR significantly different from 1, that would suggest a gender-based difference in turnover risk.

## Conclusion

In conclusion, my survival analysis indicates that the probability of employees staying decreases over time, and while there might be differences by gender, they were not strongly statistically significant in this dataset. Through the Cox regression, I identified potential risk factors—such as pay level or certain personality traits—that could influence an employee’s likelihood of leaving. These insights can guide the company’s decisions, such as adjusting compensation, offering targeted training, or tailoring retention programs, to reduce turnover and retain employees for a longer period.