SVEUČILIŠTE U RIJECI

**TEHNIČKI FAKULTET**

Diplomski studij računarstva

Projektna dokumentacija

NASLOV

Nositelj kolegija: prof. dr. sc. Zoran Čarija

Rijeka, veljača 2025. Filip Jovanović

Marko Putić

Uvod

Inicijalna ideja ovog projekta je analiza triju nepovezanih baza podataka. Tema koju obrađuje ovaj projekt je kupovna moć pojedinaca koji se bave različitim vrstama poslova te ostvaruju određeni godišnji novčani prihod. Ti se podaci namjeravaju upariti s podacima o tipovima vozila koji sadrže svoje specifikacije i cijenu. Početna faza je čišćenje podataka i formiranje baza na način da su podaci u sve 3 odabrane baze sistematizirani i da je njima moguće dobiti određene zakonitosti i rezultate.

Vrste analiza

1. Testing Relationships or Correlations
   1. Pearson Correlation Coefficient (scipy.stats.pearsonr)

Use case: Determine the linear correlation between variables, e.g., car price and sales volume.

Example: Is there a significant correlation between car age and resale value?

* 1. Spearman Rank Correlation (scipy.stats.spearmanr)

Use case: Assess monotonic relationships (non-linear correlations) between variables.

Example: Does customer income rank correlate with the type of car purchased?

1. Comparing Groups
   1. Independent t-test (scipy.stats.ttest\_ind)

Use case: Compare the means of two independent groups.

Example: Are average car prices significantly different between two regions?

* 1. Mann-Whitney U Test (scipy.stats.mannwhitneyu)

Use case: Compare medians of two groups when the data is not normally distributed.

Example: Are the median sales of electric cars higher than gas-powered cars?

* 1. ANOVA (scipy.stats.f\_oneway)

Use case: Compare means across multiple groups.

Example: Is there a significant difference in sales volume across different car brands?

* 1. Kruskal-Wallis H Test (scipy.stats.kruskal)

Use case: Non-parametric alternative to ANOVA for comparing multiple groups.

Example: Are there differences in customer satisfaction ratings across car manufacturers?

1. Distribution Analysis
   1. Chi-Square Test of Independence (scipy.stats.chi2\_contingency)

Use case: Analyze categorical data for independence.

Example: Is car ownership type (leased vs. owned) independent of geographic location?

* 1. Kolmogorov-Smirnov Test (scipy.stats.ks\_2samp)

Use case: Compare distributions of two datasets.

Example: Do the sales distributions for SUVs differ between urban and rural regions?

* 1. Shapiro-Wilk Test (scipy.stats.shapiro)

Use case: Test for normality in a dataset.

Example: Are monthly sales figures normally distributed?

1. Time Series Analysis
   1. Autocorrelation (statsmodels.tsa.stattools.acf) (not SciPy but often used alongside)

Use case: Check for repeating patterns in sales over time.

Example: Is there seasonality in monthly car sales?

1. Regression Diagnostics
   1. Levene’s Test (scipy.stats.levene)

Use case: Check for equal variance between groups before performing t-tests or ANOVA.

Example: Are variances in car sales consistent across different regions?

* 1. Durbin-Watson Test (statsmodels.stats.stattools.durbin\_watson) (not SciPy)

Use case: Check for autocorrelation in regression residuals.

Example: Are residuals from car price prediction models independent?

1. Outlier Detection
   1. Grubbs’ Test (scipy.stats.grubbs.test) (Requires external package)

Use case: Identify outliers in a dataset.

Example: Are there any unusually high sales figures in a monthly dataset?

* 1. Z-Score (scipy.stats.zscore)

Use case: Detect outliers based on standard deviation.

Example: Are there extreme values in the resale price dataset?

1. Survival Analysis
   1. Kaplan-Meier Estimator (lifelines.KaplanMeierFitter) (Requires lifelines package)

Use case: Analyze the "survival" of cars based on age or mileage.

Example: How does the probability of a car breaking down change with mileage?

Korelacije:

The choice between **Pearson**, **Spearman**, and **Kendall** correlation methods depends on the nature of your data and the assumptions you can reasonably make. Here's a breakdown of when to use each:

**1. Pearson Correlation**

* **Measures**: The strength and direction of a **linear relationship** between two continuous variables.
* **Assumptions**:
  + Both variables are **normally distributed**.
  + The relationship between the variables is **linear**.
  + There are no significant outliers, as they can heavily influence the result.
* **When to Use**:
  + You are analyzing continuous data that you believe has a linear relationship.
  + The data satisfies normality assumptions (or at least approximates normality).

**2. Spearman Correlation**

* **Measures**: The strength and direction of a **monotonic relationship** between two variables. A monotonic relationship means that as one variable increases, the other either always increases or always decreases, but not necessarily at a constant rate.
* **Assumptions**:
  + Does **not require normality**.
  + Does not require a linear relationship (can capture monotonic trends).
  + Can handle **ordinal** data (or rank-transformed data).
* **When to Use**:
  + You are analyzing data that does not meet the assumptions for Pearson correlation (e.g., it’s not linear or normally distributed).
  + You have ordinal data or data with a non-linear but monotonic relationship.
  + You need a method that is **more robust to outliers** than Pearson.

**3. Kendall Correlation**

* **Measures**: The strength and direction of a **monotonic relationship**, similar to Spearman, but based on the **ordinal ranking** of the data using pairwise concordance and discordance.
* **Assumptions**:
  + Like Spearman, it does not assume normality or linearity.
  + Can handle **ordinal data**.
  + More robust to small datasets or when you want a method based on rank concordance.
* **When to Use**:
  + You are working with **small datasets** where Spearman might give inaccurate results due to rank ties.
  + You prefer a rank-based measure that is less sensitive to data distribution.