<https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

* Analyze real-world traffic accident data using essential data science skills: data cleaning, EDA, statistical analysis, and data visualization.
* You'll demonstrate proficiency in Python, Jupyter Notebooks, and libraries like Pandas, PySpark, and Matplotlib.
* Deliverables include a portfolio-ready analysis, a Tableau dashboard, and a video presentation to showcase your ability to communicate data-driven insights.
* **Demonstrate your ability to extract actionable insights from complex datasets through:**  
  + Identifying patterns and trends in large datasets.
  + Using appropriate statistical methods to validate findings.
  + Creating clear, informative dashboard visualizations that tell a story.
  + Translating technical findings into business recommendations.
* **Build a professional portfolio piece on GitHub that:**
  + Shows your coding style and documentation practices.
  + Demonstrates your approach to problem-solving.
  + Showcases your ability to communicate technical concepts clearly.
  + Serves as a talking point in interviews.

Scenario

For the Department of Transportation (DOT), understanding and reducing traffic accidents is a critical mission that directly impacts public safety, economic costs, and quality of life across the United States.

Here's why this analysis matters from a stakeholder and business perspective:

* **Economic Impact:** Traffic accidents cost billions annually in medical expenses, property damage, and lost productivity, making even small reductions highly valuable.
* **Public Safety:** As a leading cause of injury and death, reducing traffic accidents directly fulfills DOT's core mandate to protect citizens.
* **Infrastructure Prioritization:** Data analysis enables strategic allocation of limited infrastructure improvement budgets to highest-risk areas.
* **Policy Development:** Accident data informs new safety regulations and provides metrics to evaluate existing programs' effectiveness.
* **Stakeholder Accountability:** Comprehensive analysis demonstrates evidence-based decision-making to Congress, local governments, and the public.
* **Cross-Agency Collaboration:** Shared data insights can align accident reduction efforts across DOT, law enforcement, and emergency services.
* **Technology Integration:** Understanding accident patterns guides how emerging vehicle technologies should be regulated to maximize safety benefits.

**Based on your comprehensive analysis**, you'll provide three data-driven insights that the DOT could utilize to reduce traffic accidents and improve road safety. As with any real-world data science project, you'll need to explore the dataset to determine which variables are most relevant to your analysis.

**This includes the following tasks:**

1. Formulate key questions that will guide your analysis.
2. Perform comprehensive EDA to investigate and understand the data.
3. Clean and prepare data for analysis utilizing python libraries.
4. Conduct data analysis to produce concrete findings and recommendations.
5. Support your findings with clear visualizations and statistical testing.
6. Develop an interactive dashboard that communicates your analysis results to non-technical stakeholders.

Step 1: Set up Your Project Environment

1. **Create a new GitHub repository** for the project utilizing the provided repository template.  
   * “Use this template” → “Create a new repository”.
2. **Clone repository** to your local computer.  
   * Clone into a folder you can easily access (Documents, Flatiron, Desktop).
3. **Download the US Accidents dataset** from Kaggle.
4. **Move the data file**into your cloned repository.
   * Place inside the provided Data folder.
5. **Create an environment**with any necessary libraries or use existing.

Step 2: Business Understanding

1. **Emphasize the relevancy of the data**you are using in the context of your business scenario.
2. **Create clear analytical questions** to answer via your project.
3. **Connect the why of the project** back to concrete benefits and goals.
4. **Identify clear stakeholders** for your analysis (who is your audience).

Step 3: Data Understanding

1. **Load the dataset**and **explore**its structure and basic statistics.
2. **Document the meaning**of each variable and its potential relevance to your analysis.
3. **Assess data quality issues** (missing values, outliers, inconsistencies).

Step 4: Data Prep

1. **Clean and preprocess the data** to prepare it for analysis:  
   * Handle missing values appropriately (imputation or removal).
   * Address outliers that may skew your analysis.
   * Convert data types as needed (dates, categorical variables, etc.).
   * Create derived features that might enhance your analysis.
2. **Document all data preparation steps** clearly in your notebook.
3. **Create descriptive statistics** to summarize the data.

Step 5: EDA

1. **Conduct analysis** driven by your business questions in step 2.
2. **Create different visualizations** to illustrate your findings.
3. **Include clear interpretations** for each visualization.
4. **Document** any surprising findings or counterintuitive patterns.

Step 6: Statistical Data Analysis

1. **Perform statistical tests** to validate patterns you've identified:  
   * Chi-square tests for categorical variable relationships.
   * T-tests or ANOVA for comparing group means.
   * Make sure to check assumptions and use non-parametric when appropriate.
2. **Analyze correlations** between variables.
3. **Identify significant factors**that contribute to accident severity or frequency.
4. **Use statistical measures** to support your conclusions.
5. **Document** the methodology, assumptions, and limitations of your statistical analysis.

Step 7: Insights & Recommendations

1. Based on your analysis, **identify at least three actionable recommendations** for reducing traffic accidents.
2. **Each recommendation must be**:  
   * Data-driven and supported by your analysis.
   * Specific and actionable.
   * Impactful in addressing the problem.
3. **Explain the potential impact**of each recommendation and how its effectiveness could be measured.
4. **Consider potential challenges or limitations** in implementing your recommendations.
5. **Suggest metrics** for tracking the effectiveness of your recommendations.

Step 8: Create an Interactive Dashboard

1. **Design and develop an interactive dashboard** that allows users to:  
   * Explore accident patterns across different dimensions.
   * Filter data.
   * Visualize key insights from your analysis.
   * Understand the basis for your recommendations.
2. **Ensure your dashboard is user-friendly** for non-technical stakeholders.
3. **Include appropriate explanations and context** for the visualizations.
4. **Test the dashboard**with peers to ensure clarity and usability.

Step 9: Create Deliverables

1. **Finalize your Jupyter Notebook**with well-organized well-documented code, and thorough markdown explanation of your analysis, recommendations, and process. Your notebook needs to contain a link to your Tableau dashboard.
2. **Update your GitHub repository**with all deliverables, ensure that all necessary files are provided (data and/or images)
   * Git add .
     + Make sure to Include the period.
   * Git commit -m ‘commit message here’
   * Git push origin main
3. **Ensure all code is well-commented** and follows best practices for readability and reproducibility.

**RUBRIC**

**Data Cleaning & Preparation**

* **Excelled (20):** Comprehensive data cleaning with detailed documentation. All quality issues addressed. Data transformations enhance analysis potential. Derived features add significant value.
* **Met Expectations (16):** Adequate data cleaning with documentation. Most quality issues are addressed. Appropriate transformations applied. Some useful derived features created.
* **Attempted (8):** Basic data cleaning performed. Some quality issues remain. Limited transformations. Few or no derived features.

1. **No Attempt/Incorrect (0):** Little to no data cleaning or preparation evident. Data quality issues unaddressed.

**Exploratory Data Analysis**

* **Excelled (20):** In-depth EDA with diverse, insightful visualizations. Clear patterns identified and thoroughly interpreted. Multiple angles of analysis explored. Creative approaches to uncovering hidden patterns.
* **Met Expectations (16):** Solid EDA with appropriate visualizations. Patterns identified and interpreted. Multiple aspects of data explored. Good interpretation of findings.
* **Attempted (8):** Basic EDA with simple visualizations. Limited pattern identification. Analysis lacks depth. Minimal interpretation of visualizations.
* **No Attempt/Incorrect (0):** Minimal or ineffective EDA. Few or poor-quality visualizations. No clear patterns identified.

**Statistical Analysis & Insights**

* **Excelled (20):** Sophisticated statistical analyses providing deep insights. Strong data-driven narrative. Conclusions clearly supported by evidence. Limitations acknowledged and addressed.

**Met Expectations (16):** Appropriate statistical techniques applied. Good insights derived. Conclusions supported by data. Some acknowledgment of limitations.

* **Attempted (8):** Basic statistical analysis. Limited insights. Some conclusions lack strong evidence. Limitations not well-addressed.
* **No Attempt/Incorrect (0)** Minimal or incorrect statistical analysis. Few valid insights. Conclusions unsupported by data.

**Interactive Dashboard**

* **Excelled (20):** Exceptional dashboard with intuitive design, multiple interactive elements, and clear insights. Appropriate for non-technical users. Enhances understanding of complex patterns.
* **Met Expectations (16):** Functional dashboard with good interactive features and clear presentation of key findings. Accessible to non-technical users.
* **Attempted (8):** Basic dashboard with limited interactivity or clarity. May be challenging for non-technical users.
* **No Attempt/Incorrect (0):** Missing, non-functional, or ineffective dashboard.

**Recommendations & Business Value**

* **Excelled (20):** Three excellent, data-driven recommendations with clear implementation paths and impact measurements. Exceptional business value demonstrated. Recommendations address root causes.
* **Met Expectations (16):** Three solid, data-driven recommendations with implementation considerations and potential impact described. Good business value. Recommendations address important factors.
* **Attempted (8):** Fewer than three recommendations or recommendations with limited data support. Vague implementation details. Limited business value addressed.
* **No Attempt/Incorrect (0):** Missing or ineffective recommendations. No clear business value. Recommendations not supported by analysis.