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

https://github.com/learn-co-curriculum/DS\_Capstone\_Template#

<https://highways.dot.gov/safety>

https://www.nhtsa.gov/research

This project will assess your ability to perform comprehensive data analysis on real-world traffic accident data. You'll apply fundamental data science skills including **data cleaning, exploratory data analysis (EDA), statistical analysis, data visualization**, and **deriving business insights**.

Through this project, you'll demonstrate your proficiency in **Python, Jupyter Notebooks**, and **data analysis libraries** like Pandas, Pyspark, and Matplotlib, all while developing a portfolio-ready project that showcases your analytical abilities. In addition, you will also produce a Tableau dashboard and video presentation to demonstrate your communication skills in relation to data science.

Methodological Requirements

* **Effective Visualization:** Create clear, compelling visualizations that effectively communicate your findings. Remember that well-designed graphics can reveal patterns not immediately obvious in raw data.
* **Statistical Rigor:** Apply appropriate statistical tests to validate relationships between variables. Your analysis should distinguish between correlation and causation where possible.
* **Transparent Methodology:** Document your approach clearly, including data cleaning steps, analytical choices, and limitations. This ensures your work is reproducible and credible.
* **Accessible Insights:** Frame your findings in ways that non-statistical audiences can understand. The most valuable insights are those that can be easily communicated to stakeholders and decision-makers.

Scenario

You've been hired as a data analyst for the Department of Transportation (DOT). The department is concerned about the number of traffic accidents across the United States and wants to develop strategies to reduce accidents and improve road safety.

**Your task:** Analyze the US Accidents dataset to identify patterns, trends, and factors contributing to accidents. 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.

**Potential Areas of Analysis**

* **Spatial and Temporal Patterns:**Consider examining when and where accidents most frequently occur. You might explore patterns by time of day, day of week, season, and geographic location. This type of analysis could potentially reveal critical hotspots and time periods requiring intervention.
* **Environmental Factors:** You could investigate how weather conditions correlate with accident rates. Consider analyzing how visibility, precipitation, temperature, and other environmental variables might affect driver behavior and road conditions.
* **Infrastructure Considerations:** One possible avenue is to identify specific road features associated with accident severity. This might include road design, signage, lighting, or other infrastructural elements that could contribute to or mitigate accident risk.
* **Urban vs. Rural Comparison:** You may want to compare accident patterns between urban and rural settings. These different environments likely present distinct challenges and risk factors that might require tailored safety approaches.

Problem Solving Process

To successfully complete this project, follow the **CRISP-DM**(Cross-Industry Standard Process for Data Mining) methodology:

1. **Business Understanding:** Define the problem clearly and identify key questions that your analysis should answer. Determine how success will be measured and what insights would be valuable to stakeholders.
2. **Data Understanding:** Explore the US Accidents dataset to understand its structure, variables, scope, and limitations. Identify potential issues with the data and assess its suitability for addressing the business problem.
3. **Data Preparation:** Clean and preprocess the data, handling missing values, outliers, and duplicates. Create derived features that might provide additional insights (e.g., time of day categories, weather condition groupings).
4. **Analysis/Statistical Analysis:** (Instead of Modeling in traditional CRISP-DM) Apply appropriate statistical methods to identify patterns, correlations, and significant factors. Test hypotheses about accident causes and contributing factors.
5. **Evaluation:** Interpret your findings in the context of the business problem. Assess whether your analysis provides actionable insights and addresses the key questions identified in the business understanding phase.
6. **Deployment/Communication:** (Dashboard + Notebook) Create an interactive dashboard for stakeholders to explore your findings. Develop a well-documented Jupyter notebook that explains your process and conclusions. Prepare a presentation that effectively communicates your insights and recommendations.

Following this standardized process ensures a methodical approach to your analysis, helps maintain focus on business objectives, and results in a comprehensive solution that addresses all aspects of the problem.

* 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.

**Potential Areas of Analysis**

* **Spatial and Temporal Patterns:**Consider examining when and where accidents most frequently occur. You might explore patterns by time of day, day of week, season, and geographic location. This type of analysis could potentially reveal critical hotspots and time periods requiring intervention.
* **Environmental Factors:** You could investigate how weather conditions correlate with accident rates. Consider analyzing how visibility, precipitation, temperature, and other environmental variables might affect driver behavior and road conditions.
* **Infrastructure Considerations:** One possible avenue is to identify specific road features associated with accident severity. This might include road design, signage, lighting, or other infrastructural elements that could contribute to or mitigate accident risk.
* **Urban vs. Rural Comparison:** You may want to compare accident patterns between urban and rural settings. These different environments likely present distinct challenges and risk factors that might require tailored safety approaches.

Problem Solving Process

To successfully complete this project, follow the **CRISP-DM**(Cross-Industry Standard Process for Data Mining) methodology:

1. **Business Understanding:** Define the problem clearly and identify key questions that your analysis should answer. Determine how success will be measured and what insights would be valuable to stakeholders.
2. **Data Understanding:** Explore the US Accidents dataset to understand its structure, variables, scope, and limitations. Identify potential issues with the data and assess its suitability for addressing the business problem.
3. **Data Preparation:** Clean and preprocess the data, handling missing values, outliers, and duplicates. Create derived features that might provide additional insights (e.g., time of day categories, weather condition groupings).
4. **Analysis/Statistical Analysis:** (Instead of Modeling in traditional CRISP-DM) Apply appropriate statistical methods to identify patterns, correlations, and significant factors. Test hypotheses about accident causes and contributing factors.
5. **Evaluation:** Interpret your findings in the context of the business problem. Assess whether your analysis provides actionable insights and addresses the key questions identified in the business understanding phase.
6. **Deployment/Communication:** (Dashboard + Notebook) Create an interactive dashboard for stakeholders to explore your findings. Develop a well-documented Jupyter notebook that explains your process and conclusions. Prepare a presentation that effectively communicates your insights and recommendations.

Following this standardized process ensures a methodical approach to your analysis, helps maintain focus on business objectives, and results in a comprehensive solution that addresses all aspects of the problem.

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.