ANALYTICS ALCHEMISTS

By / Analytics Alchemists

Digital Egypt Pioneers

ROAD ACCIDENT ANALYSIS

Road Accidents: Causes, Patterns, and Prevention

ANALYTICS ALCHEMISTS

OUR TEAM

Digital Egypt Pioneers

- Sayed Elmasry
- Dai Guevara
- Yomna Ahmed
- Mohamed Radi
- Nadine Elkemery

ANALYTICS ALCHEMISTS

Digital Egypt Pioneers

Introduction

Road accidents are a serious issue affecting thousands of lives every year in the UK.
Through data analysis, we aimed to understand when, where, and why these accidents happen — and what actions can help reduce them.
This project uses real accident data to uncover patterns, identify risk factors, and support smarter decisions for safer roads.

ANALYTICS ALCHEMISTS

Overview about data

Digital Egypt Pioneers

- This data covers incidents for the years (2020-2021).

🔍 2. Key Features

- Date & Time of accidents
- Location & road type
- Severity (Slight / Serious / Fatal)
- Weather, lighting & road surface conditions
- Number of vehicles and casualties

1. General Info

- Real UK road accident dataset
- Covers over 307K records
- Time range: multiple years
- Data source: official government records

ANALYTICS ALCHEMISTS

Workflow

- Data cleaning & preprocessing
- Feature Selection / Engineering
- SQL & Exploratory Data Analysis (EDA)
- Modeling
- Interpretation & Visualization
- Reporting

Digital Egypt Pioneers

ANALYTICS ALCHEMISTS

Cleaning & processing

Digital Egypt Pioneers

Data Cleaning

- Convert date and time columns.
- Remove duplicates.
- Sort data by date.
- Handle missing values in multiple ways
- Fill with most frequent values.
- Forward fill.
- Custom fill by season.

Outliers Handling

- Reduce large values in the number of infected and number of vehicles to the median if they exceed a certain threshold.
- Display graphs showing the distribution of values after adjustment.

Simplify columns

- vehicle types into general categories (e.g., Car, Motorcycle, Bus, etc.).
- weather conditions into simpler categories (e.g., Clear, Rainy, Snowy, etc.).
- light condition into simpler categories (e.g., No Lighting ,Lights Lit, etc.).

ANALYTICS ALCHEMISTS

CHECKING FOR NULLS

Digital Egypt Pioneers

```
# Looking for missing values
print(df.isnull().sum())
```

The second image shows the result of running this code. As I explained in the previous message, there are several columns with different numbers of missing values, which I decided to handle using more than one method.

```
most_common_road = df['Road_Type'].mode()[0]
df['Road_Type'] = df['Road_Type'].fillna(most_common_road)

print("num of nulls :", df['Road_Type'].isnull().sum())
```

```
num of nulls : 0
```

```
df['Weather_Conditions'] = df['Weather_Conditions'].fillna(method='ffill')
print("num of nulls :", df['Weather_Conditions'].isnull().sum())
```

```
num of nulls : 0
```

```
Accident_Index                  0
Accident Date                   0
Month                           0
Day_of_Week                     0
Year                            0
Junction_Control                0
Junction_Detail                 0
Accident_Severity               0
Latitude                        0
Light_Conditions                0
Local_Authority_(District)      0
Carriageway_Hazards        302549
Longitude                       0
Number_of_Casualties            0
Number_of_Vehicles              0
Police_Force                    0
Road_Surface_Conditions       317
Road_Type                    1534
Speed_limit                     0
Time                           17
Urban_or_Rural_Area             0
Weather_Conditions           6057
Vehicle_Type                    0
dtype: int64
```

ANALYTICS ALCHEMISTS

HANDLE OUTLIERS

Digital Egypt Pioneers

```python
threshold = 10  # أقصى عدد مصابين نعتبره طبيعي
median = df['Number_of_Casualties'].median()  # 1 = اللي هو غالباً

# استبدال أي رقم أكبر من threshold بالوسيط
df['Number_of_Casualties'] = df['Number_of_Casualties'].apply(
    lambda x: median if x > threshold else x
)
```

>>>>>

```python
df['Number_of_Casualties'].value_counts().sort_index()
```

```
Number_of_Casualties
1.0     235167
2.0      50088
3.0      14338
4.0       5348
5.0       1896
6.0        713
7.0        242
8.0        102
9.0         44
10.0        34
Name: count, dtype: int64
```

The value 1.0 is the most frequent (235,167 times), supporting that the median is often 1.
Values from 1 to 10 are present, but with fewer occurrences as the number increases.
There are no values greater than 10, demonstrating the code's success in replacing extreme values (greater than 10) with the median.

ANALYTICS ALCHEMISTS

Feature Selection & Engineering

Digital Egypt Pioneers

To improve the quality of the analysis and extract deeper insights from the data, several new columns were created based on the original data. These derived features help better understand accident patterns by time, season, and road conditions.

Weekend

Add a column specifying whether the incident occurred on a weekend.

Peak Hours

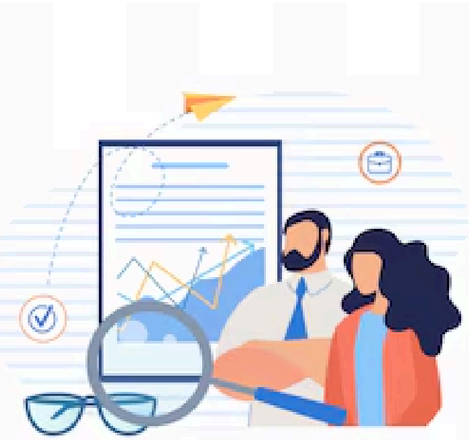
Specify whether the incident occurred during peak hours.

Time Slot

Divide time into periods (morning, evening, etc.).

Season

Extract the season from the incident date.



ANALYTICS ALCHEMISTS

FEATURE ENGINEERING

Digital Egypt Pioneers

- add more features to data to improve the quality of analysis

```python
# Create Weekend column
df['Is_Weekend'] = df['Accident Date'].dt.dayofweek >= 5  # 5 = Saturday, 6 = Sunday
```

```python
# Create Peak hour column

def is_peak_hour(time):
    # تحويل الوقت إلى ساعة
    hour = time.hour
    # 7:00 ساعات الذروة من AM 9:00 إلى AM 4:00 ومن PM 6:00 إلى PM
    if (7 <= hour < 9) or (16 <= hour < 18):
        return True
    else:
        return False

# إضافة عمود 'is_peak_hour' بناءً على وقت الحادث
df['is_peak_hour'] = df['Time'].apply(is_peak_hour)
```

```python
# دالة لتصنيف الساعة إلى فترات زمنية في اليوم:

def get_time_slot(hour):
    if 0 <= hour < 6:            # - Late Night: من منتصف الليل حتى 6 صباحًا
        return 'Late Night'
    elif 6 <= hour < 12:         # - Morning: من 6 صباحًا حتى 12 ظهرًا
        return 'Morning'
    elif 12 <= hour < 18:        # - Afternoon: من 12 ظهرًا حتى 6 مساءً
        return 'Afternoon'
    else:                        # - Evening: من 6 مساءً حتى منتصف الليل
        return 'Evening'

# - Evening: من 6 مساءً حتى منتصف الليل
df['Time Slot'] = df['Time'].dt.hour.apply(get_time_slot)
```

ANALYTICS ALCHEMISTS

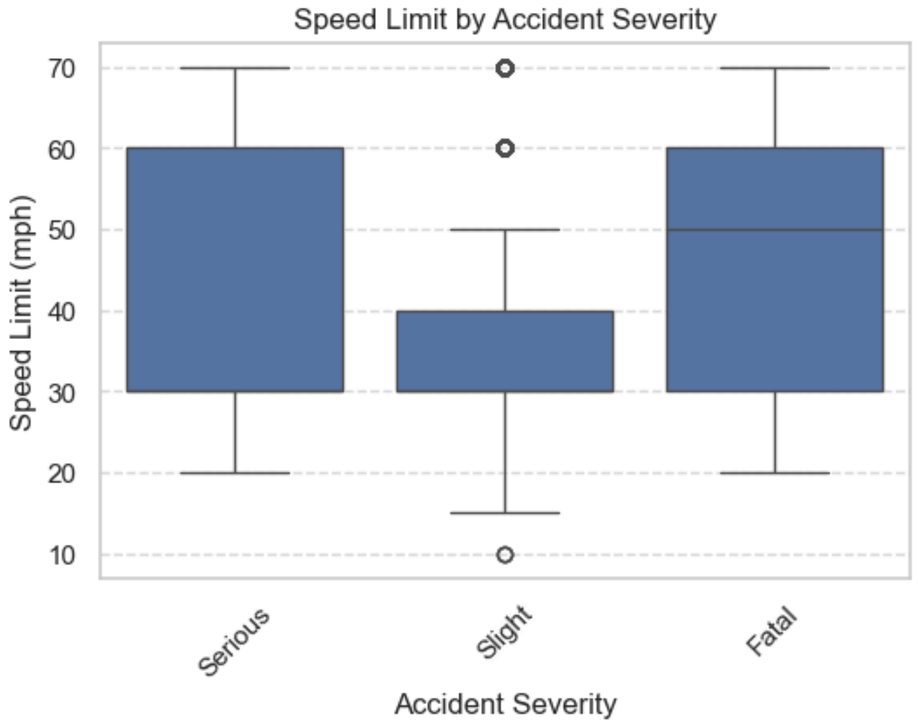
Exploratory Data Analysis (EDA)

Digital Egypt Pioneers

This exploratory data analysis aims to uncover key patterns and trends in road accident data. The goal is to better understand the factors influencing accident severity and provide insights that could support decision-making and road safety improvements. We used statistical and visual techniques to explore the dataset and highlight meaningful relationships.

Speed Limit

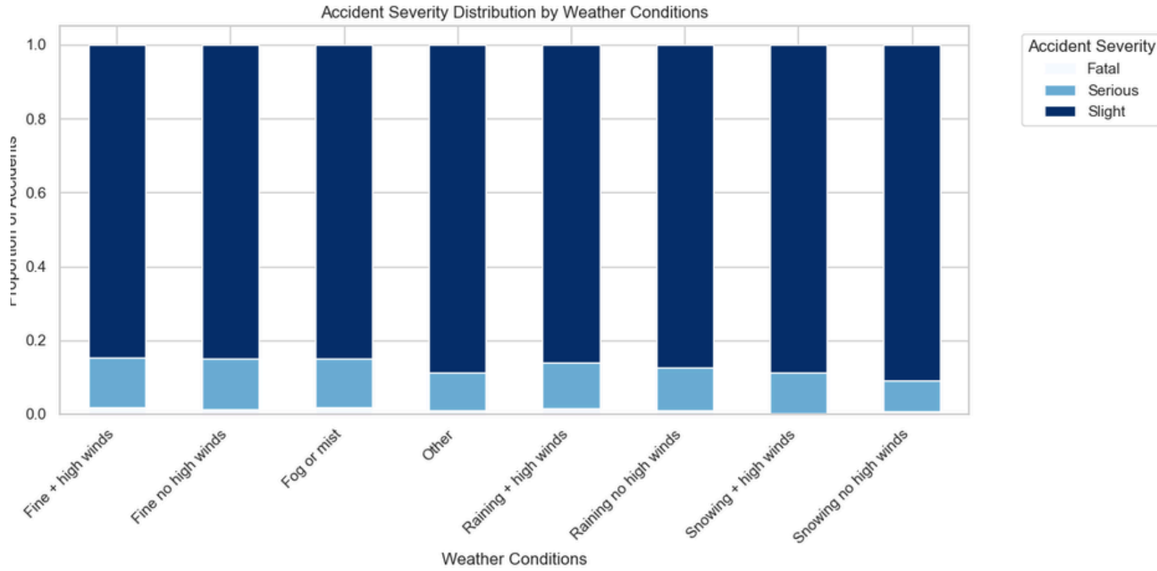
- Higher speed limits are associated with more severe accidents.
- Minor accidents often occur in areas with lower speed limits.
- Overlapping speed limits indicate other factors are at play.
- Speed is not the only factor in the severity of an accident.



ANALYTICS ALCHEMISTS

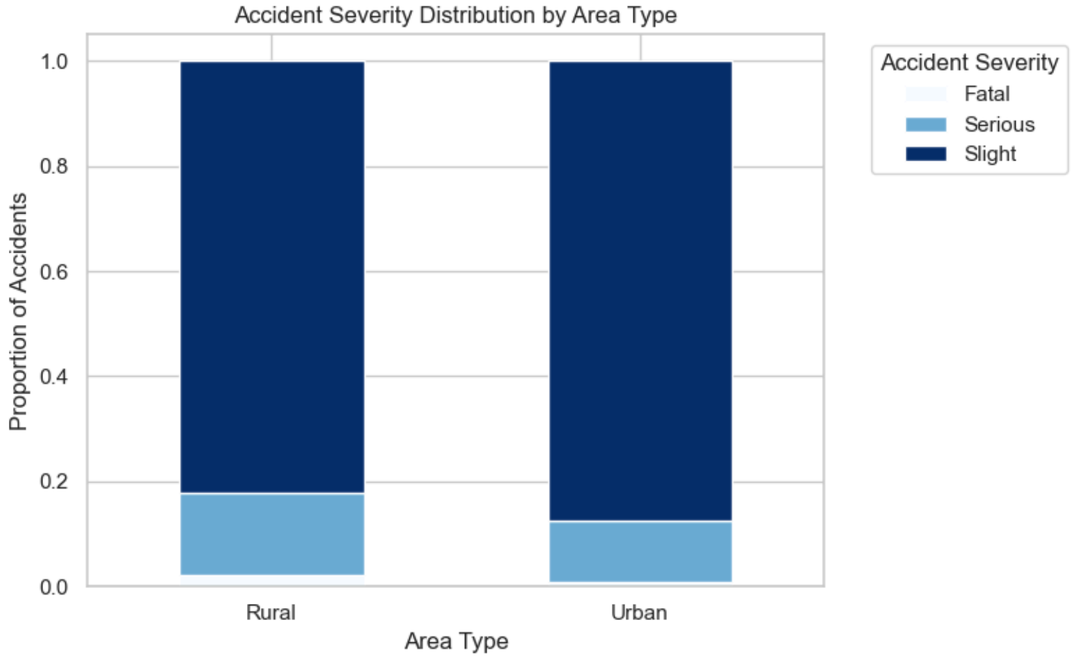
Weather Conditions

- The majority of accidents are minor in all circumstances.
- The percentage of serious and fatal accidents is low.
- The severity distribution is similar regardless of weather.
- There is no clear effect of weather conditions on the severity of accidents in this data.



Area Type

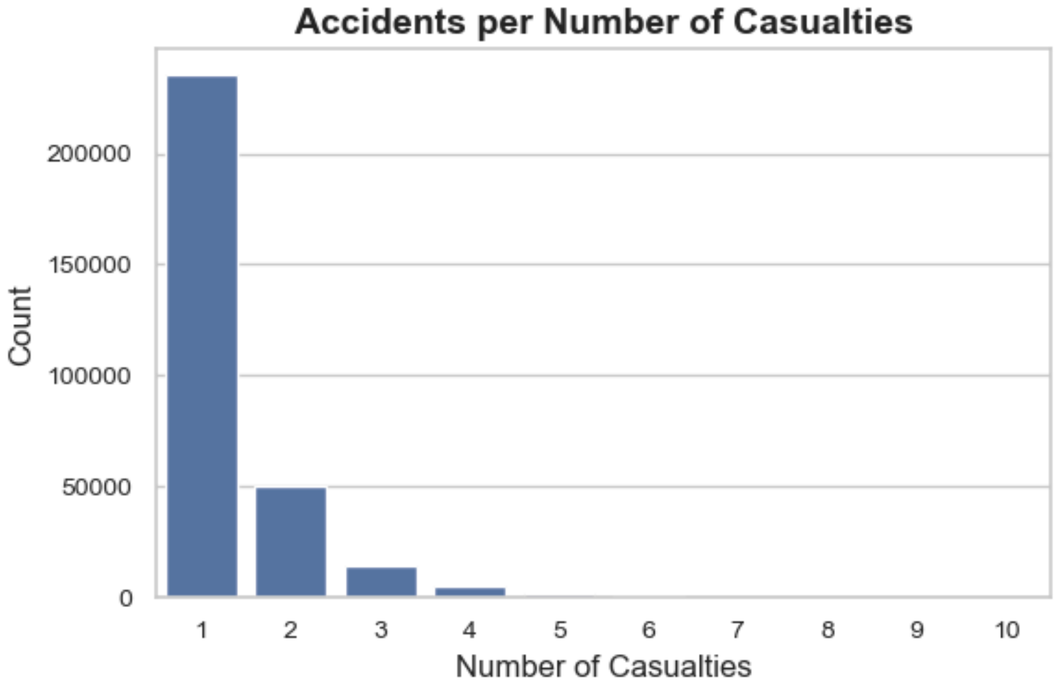
- Most accidents were slight in both rural and urban areas.
- Serious accidents are more common in rural areas than in urban areas.
- Fatal accidents are present, but very rare in both areas.



ANALYTICS ALCHEMISTS

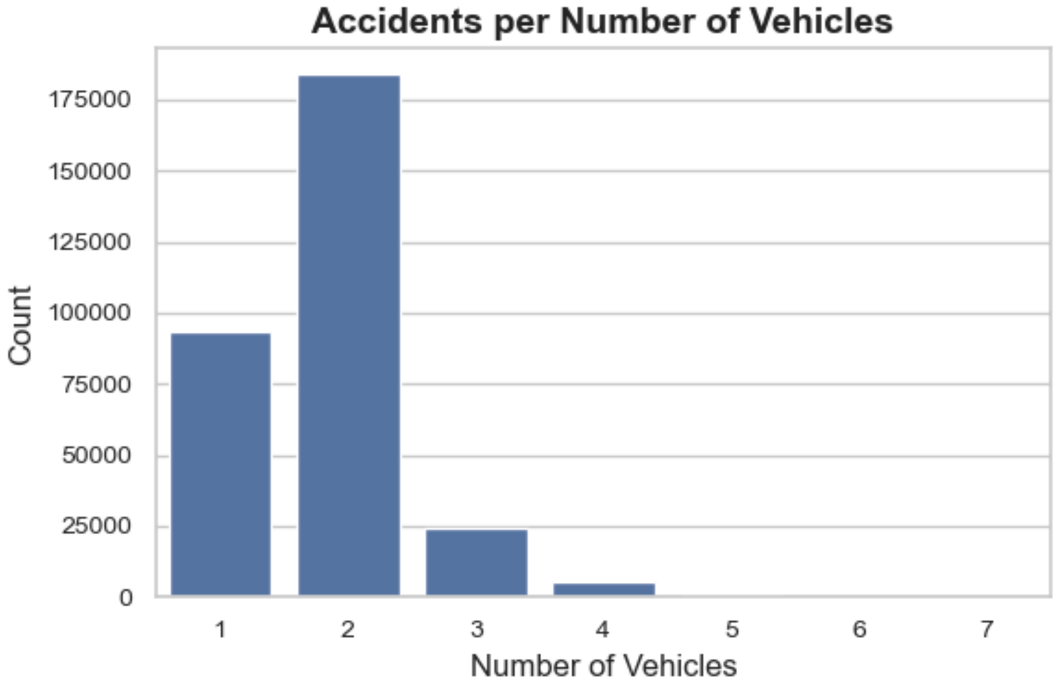
Number of casualties

- Most accidents involve only one victim, with over 225,000 cases.
- As the number of victims increases, the number of accidents decreases significantly.
- Accidents involving four or more victims are extremely rare.
- Accidents involving five to ten victims are virtually nonexistent compared to other categories.



Number of vechicles

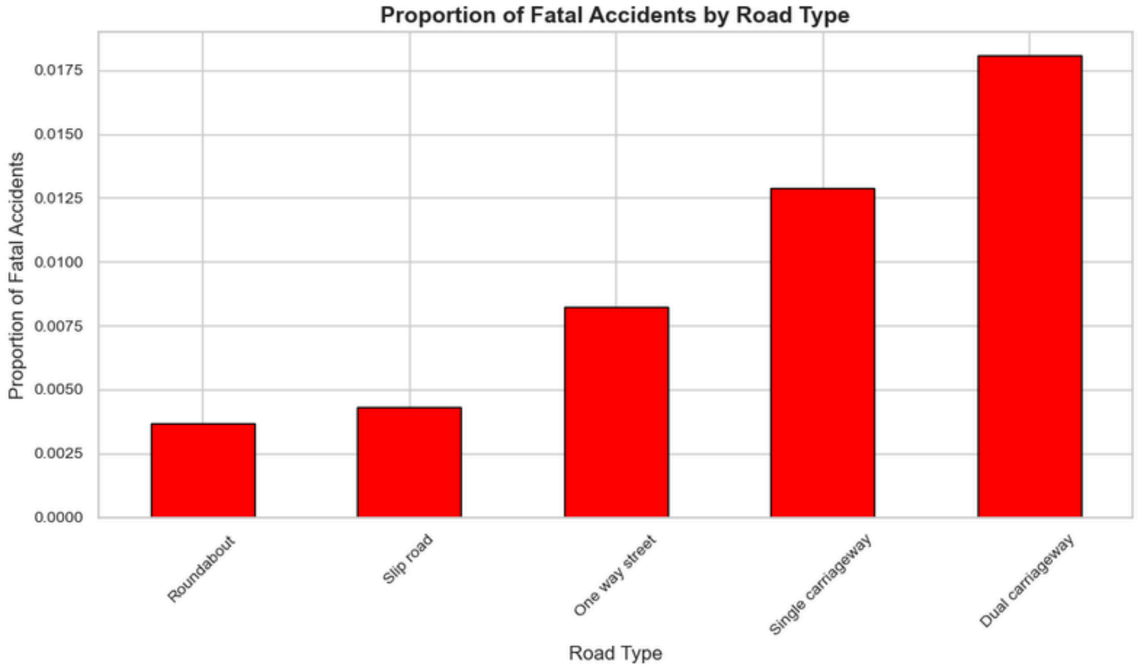
- Most accidents occur when only two vehicles are involved in a collision.
- Furthermore, single-vehicle accidents are also common.
- As the number of vehicles involved in an accident increases, the number of accidents decreases.
- This means that accidents involving three or four vehicles are less common, while those involving five or more are extremely rare.



ANALYTICS ALCHEMISTS

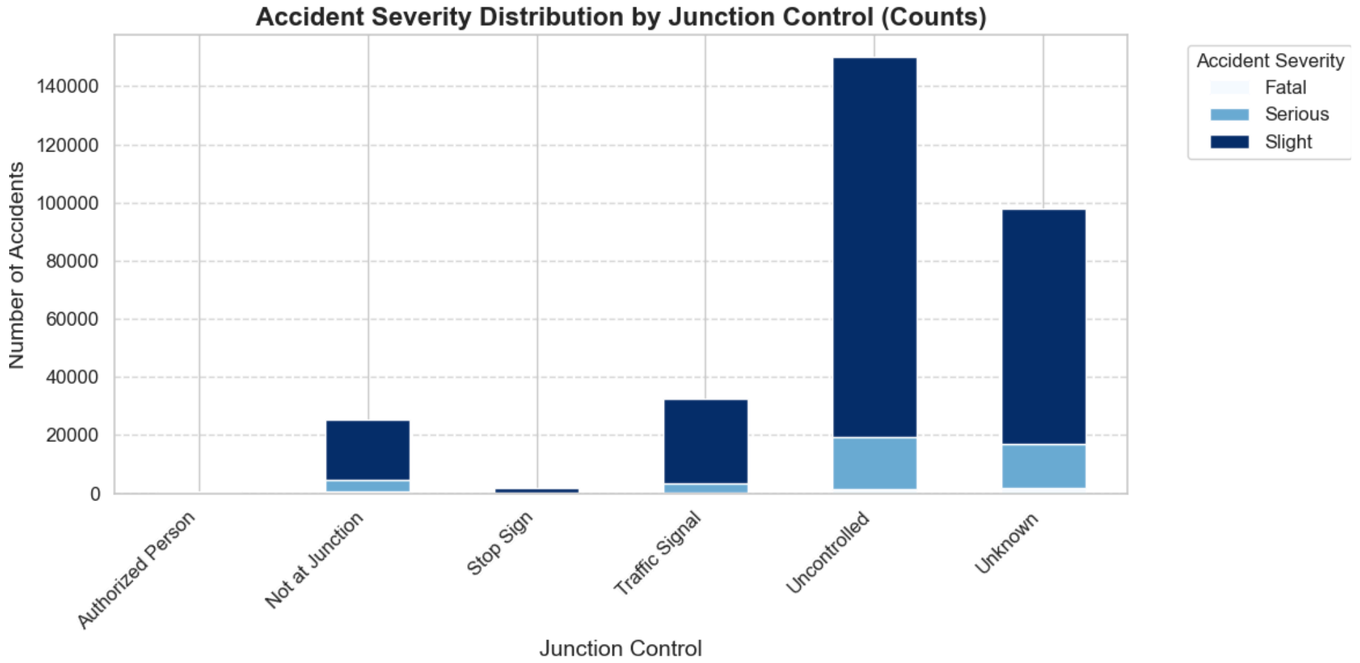
Road Type

- Roundabouts are safer due to lower speeds and traffic regulation.
- Slip-through and one-way roads have slightly higher fatality rates due to speed changes or increased collision probability.
- Single-way roads show a significant increase in fatal accidents due to the absence of a center divider.
- Dual-way roads have the highest fatality rate, likely due to higher speeds.



Junction Control

- Uncontrolled intersections recorded the highest number of accidents, particularly minor and serious ones.
- Traffic signals ranked second in the number of accidents, most of which were also minor.
- Locations outside of intersections recorded a significant number of minor accidents.
- Stop signs and authorized personnel recorded the lowest number of accidents overall.
- Unknown control was associated with a significant number of accidents, especially minor ones.



ANALYTICS ALCHEMISTS

SQL Analysis

Digital Egypt Pioneers

Analysis of Accidents by Time and Day

- Most accidents occur in the Afternoon, especially on Fridays, followed by Tuesdays and Wednesdays.
- Slight accidents are most common overall, with a clear peak on Friday Afternoon.
- Serious accidents also peak in the Afternoon, particularly on Fridays and weekends.
- Fatal accidents, though much fewer, are concentrated in the Afternoon and Late Night, especially on Fridays and Saturdays.
- Late Night has fewer accidents overall, but a higher proportion of fatal cases, suggesting risks like fatigue or impaired driving.

	Time_Slot	Day_of_Week	Accident_Severity	Accident_Count
1	Afternoon	Friday	Slight	19475
2	Afternoon	Tuesday	Slight	17278
3	Afternoon	Wednesday	Slight	17172
4	Afternoon	Thursday	Slight	16982
5	Afternoon	Monday	Slight	16668
6	Afternoon	Saturday	Slight	14754
7	Morning	Tuesday	Slight	12806
8	Morning	Wednesday	Slight	12783
9	Morning	Monday	Slight	12218
10	Afternoon	Sunday	Slight	11963
11	Morning	Thursday	Slight	11895
12	Morning	Friday	Slight	11738
13	Evening	Friday	Slight	10692
14	Evening	Wednesday	Slight	9028
15	Evening	Thursday	Slight	8981

Key Insight:

Time of day and day of week significantly affect accident severity — Fridays and weekends during Afternoon and Late Night are the most critical periods.

ANALYTICS
ALCHEMISTS

SQL Analysis

- The number of accidents saw significant fluctuations month by month during 2021 and 2022.
- The largest decrease in the number of accidents occurred in January 2022, at -27.3%, followed by December 2022, at -29.34%.
- Conversely, there was a significant increase in March 2021, at +20.57%, and in March 2022, at +12.86%.
- The first and last months of each year typically experience a decrease, which may reflect the impact of winter weather or driving conditions during those periods.

	Month	TotalAccidents	PercentageChange
1	2021-01	13416	0%
2	2021-02	10950	-18.38%
3	2021-03	13202	20.57%
4	2021-04	12715	-3.69%
5	2021-05	13811	8.62%
6	2021-06	13936	0.91%
7	2021-07	14300	2.61%
8	2021-08	13415	-6.19%
9	2021-09	13792	2.81%
10	2021-10	14834	7.56%
11	2021-11	15473	4.31%
12	2021-12	13709	-11.4%
13	2022-01	9967	-27.3%
14	2022-02	10935	9.71%
15	2022-03	12341	12.86%
16	2022-04	11510	-6.73%
17	2022-05	12372	7.49%
18	2022-06	12812	3.56%
19	2022-07	12653	-1.24%
20	2022-08	12088	-4.47%
21	2022-09	12960	7.21%
22	2022-10	13534	4.43%
23	2022-11	13622	0.65%
24	2022-12	9625	-29.34%

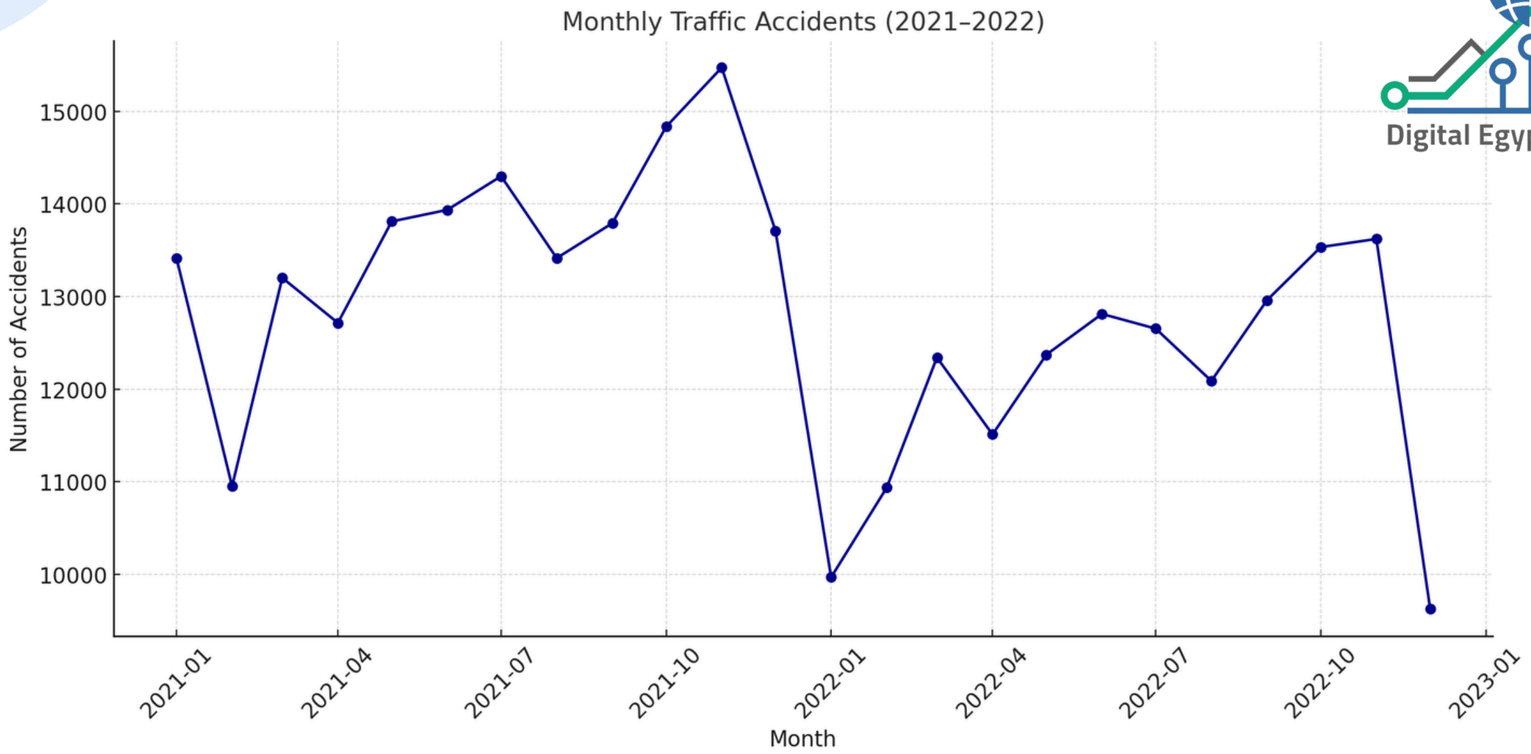
Key Insight:

Monthly changes in the number of accidents indicate the impact of seasonal factors and weather conditions on traffic safety.

ANALYTICS ALCHEMISTS

SQL Analysis

Digital Egypt Pioneers



- 🔺 November 2021 = Highest accidents
- 🔻 December 2022 = Lowest accidents
- 📉 Early 2022 = Noticeable decline
- 🍂 Fall = Increase in accidents
- ❄️ Winter = Noticeable decline

- The highest number of accidents was in November 2021, when the number of accidents exceeded 15,000.
- The lowest number of accidents was recorded in December 2022, with a sharp decline.
- There was a noticeable decline in January 2022 compared to December 2021, indicating a lower start to the year in terms of accidents.
- In general, there is a clear seasonal trend, with the number of accidents rising in the fall months (September-November) and declining at the beginning of the year (January-February).
- 2022 saw relatively less fluctuation in the number of accidents compared to 2021, but with a greater decline towards the end of the year.

Dashboard

UK Road Accident Dashboard

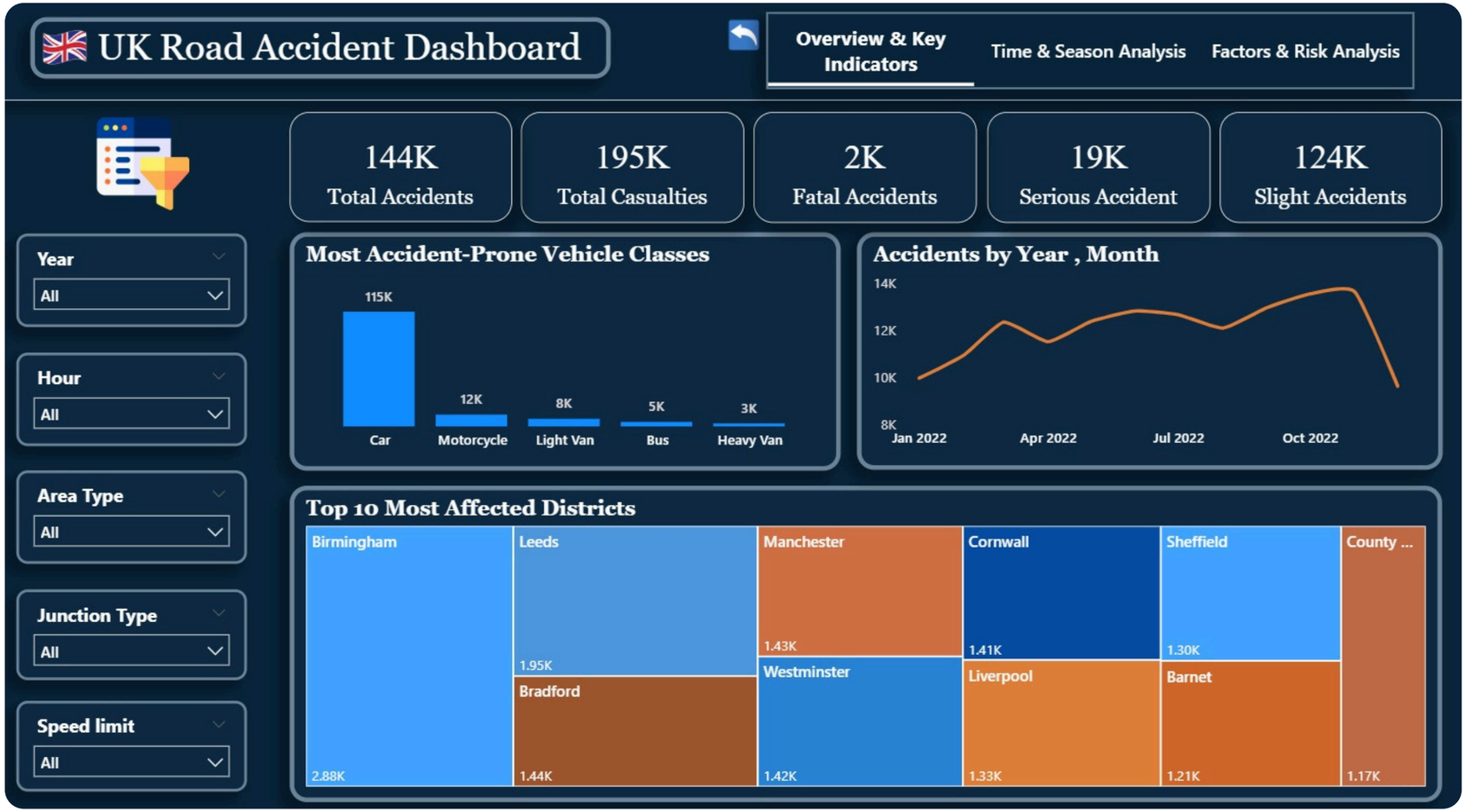
Yearly trend of accidents across the dataset

Most common vehicle types involved in accidents.

Top 10 districts with the highest accident counts.

This dashboard provides a visual summary of accident patterns based on road type and severity.

Digital Egypt Pioneers

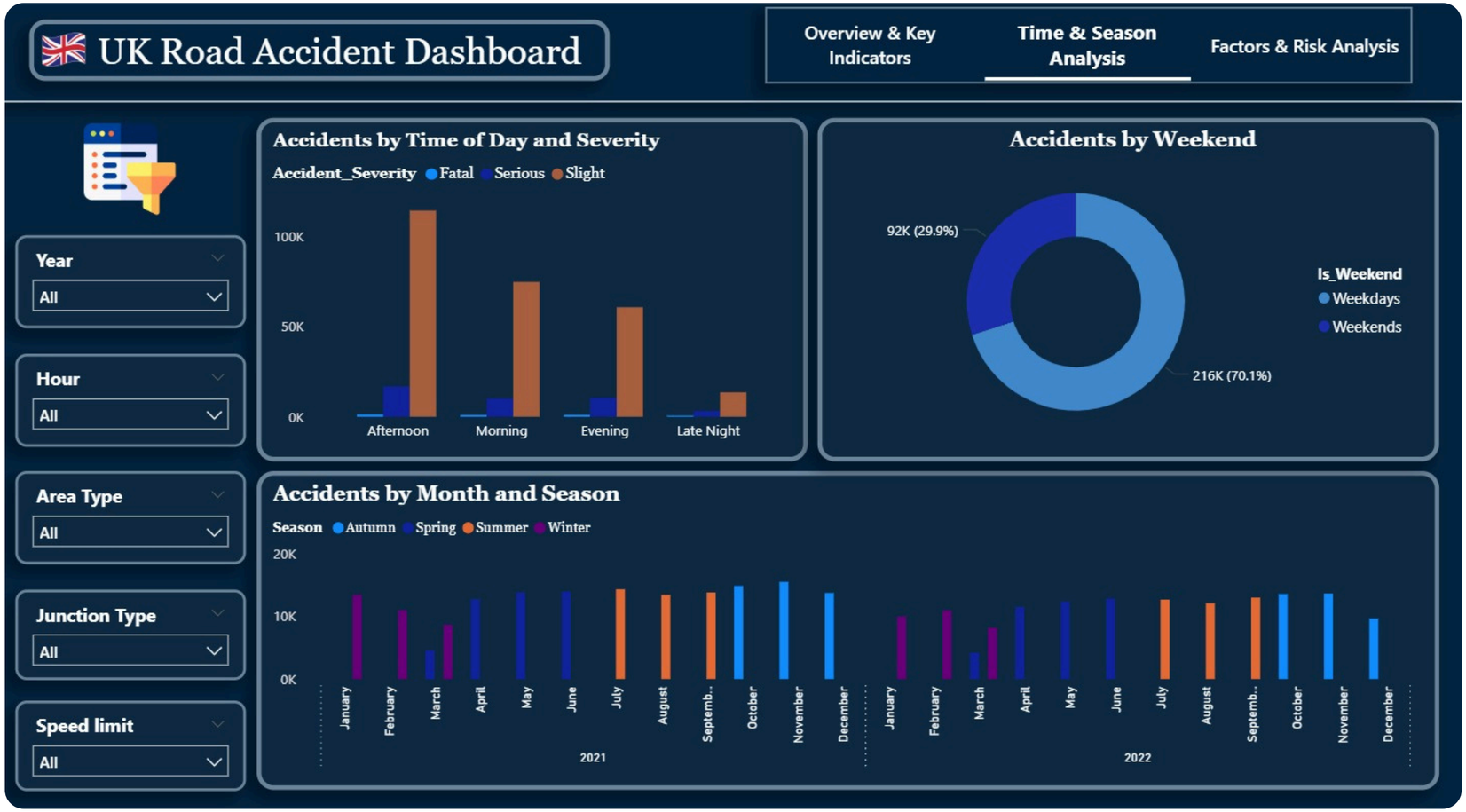


Dashboard

UK Road Accident Dashboard

This dashboard provides a visual summary of accident patterns based on road type and severity.

- Accident distribution by time of day and severity level
- Comparing accident rates on weekends vs weekdays
- Monthly and seasonal trends in accident occurrences.



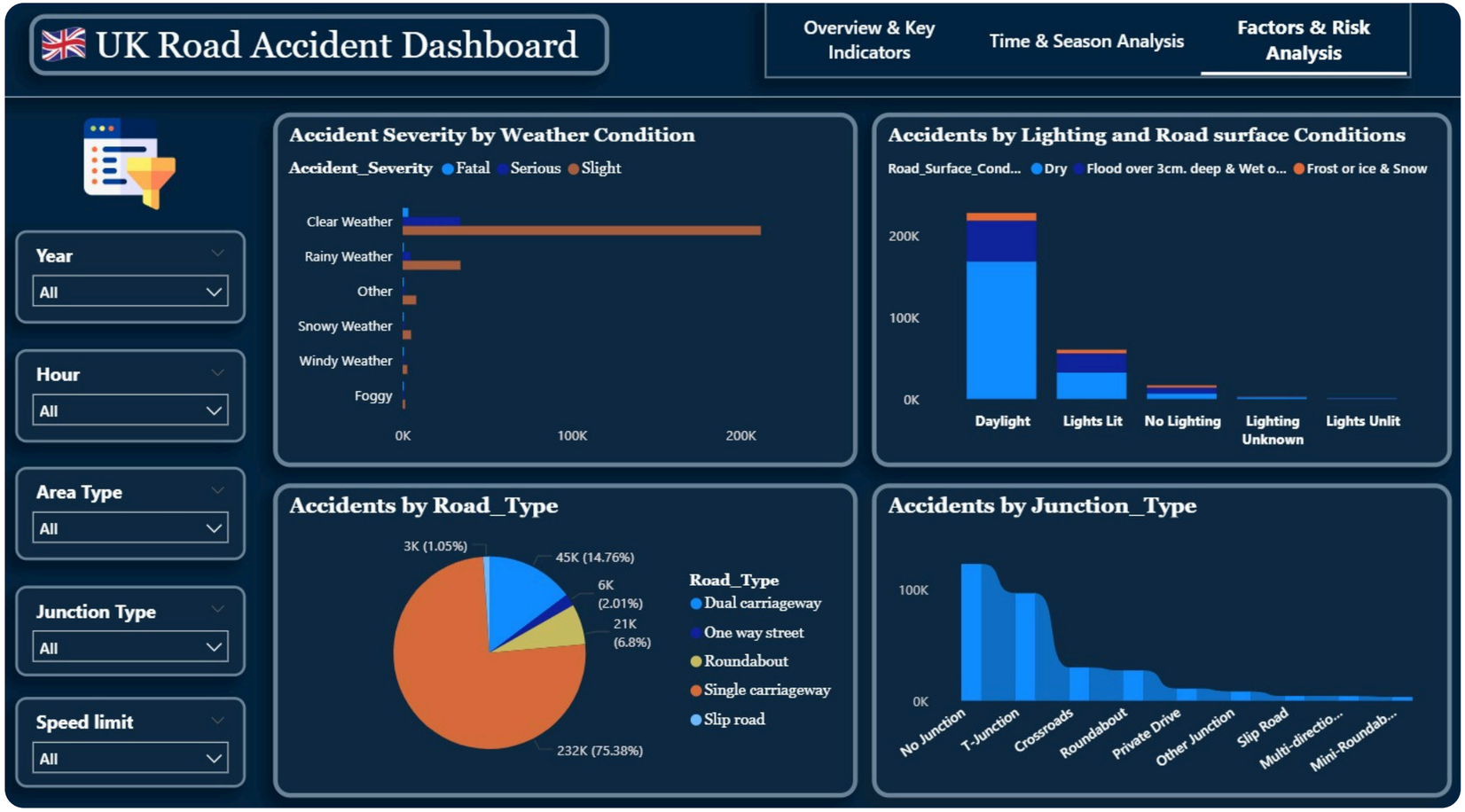
Dashboard

UK Road Accident Dashboard

- Impact of weather conditions on accident severity.
- Distribution of accidents by road type.
- Analysis of accidents by junction type.
- Effect of lighting and road surface conditions on accidents.

This dashboard provides a visual summary of accident patterns based on road type and severity.

Digital Egypt Pioneers



ANALYTICS ALCHEMISTS

Predictive Modeling

Digital Egypt Pioneers

We applied various machine learning models to predict road accidents and identify key factors. The goal was to select the best model based on metrics like accuracy, recall, and precision for reliable predictions

Logistic Regression

Applied as a baseline model for classification due to its simplicity and interpretability.

Random Forest

Chosen for its high accuracy and ability to handle multi-class classification.

Gaussian Naive Bayes

Interpretable model with fast training and predection, used in text classification

ANALYTICS ALCHEMISTS

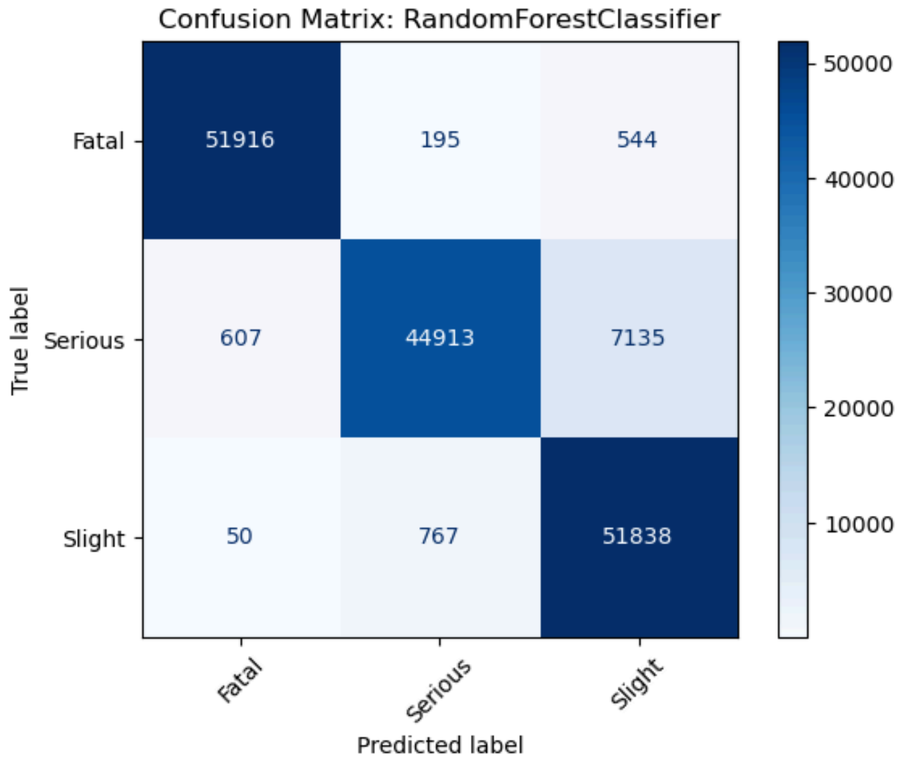
Random forset

Digital Egypt Pioneers

Slight (Class 0) and Fatal (Class 2) categories were predicted with high accuracy, with most instances correctly classified.

Some confusion was observed between Serious (Class 1) and Fatal, indicating potential overlap in their features.

Overall, the model demonstrates reliable classification ability and can support decision-making in accident analysis.




```
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=best_model.classes_)
disp.plot(cmap='Blues', xticks_rotation=45)
plt.title(f'Confusion Matrix: {best_model_name}')
plt.show()
```

ANALYTICS ALCHEMISTS

Predection

Digital Egypt Pioneers

	Actual	Predicted
95105	Slight	Slight
421522	Fatal	Fatal
603511	Serious	Serious
476529	Fatal	Fatal
556027	Fatal	Serious
739553	Serious	Serious
247354	Slight	Slight
66855	Slight	Slight
765928	Serious	Serious
763623	Serious	Serious
125379	Slight	Slight
604046	Serious	Serious
27418	Slight	Slight
558003	Fatal	Fatal
490248	Fatal	Fatal
333249	Fatal	Fatal
287232	Slight	Slight
739857	Serious	Serious
119919	Slight	Slight
567538	Serious	Serious

This table displays the first 20 rows of the model's results, and shows a direct comparison between the actual values and the model's predicted values for each event.

```
results_df = pd.DataFrame({
    'Actual': y_test,
    'Predicted': y_pred
})
results_df.head(20)  # عرض أول 20 صف
```

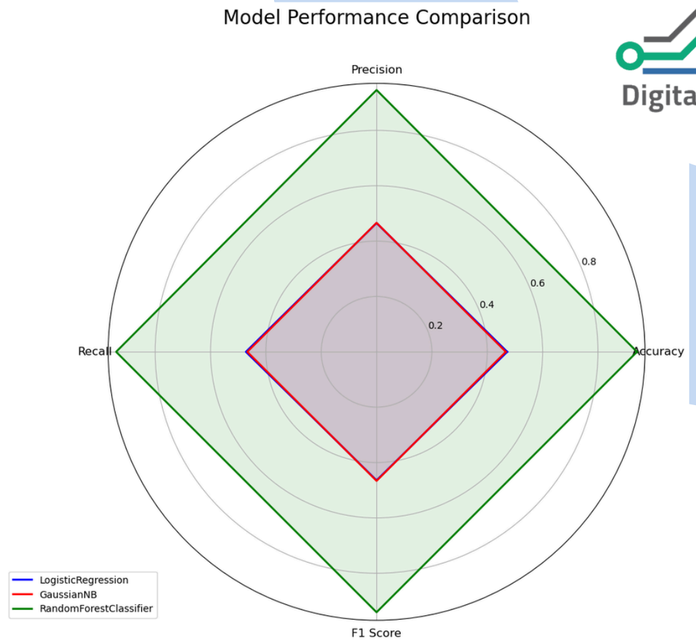
- **Sample Predictions Overview**

  - The table shows a comparison between the actual values and the predictions for the first 20 cases in the test set.
  - Most of the predictions were correct, reflecting the model's good accuracy.
  - There were some errors, such as a case predicted as "Serious" that was actually "Fatal," illustrating potential overlap between the two categories.

ANALYTICS ALCHEMISTS

Performance

```python
def radar_chart(df, title):
    categories = list(df.columns)
    N = len(categories)
    models = list(df.index)
    angles = [n / float(N) * 2 * np.pi for n in range(N)]
    angles += angles[:1]

    fig, ax = plt.subplots(figsize=(10, 10), subplot_kw=dict(polar=True))
    plt.xticks(angles[:-1], categories, size=12)
    colors = ['b', 'r', 'g', 'y', 'c', 'm']
    for i, model_name in enumerate(models):
        values = df.loc[model_name].values.tolist()
        values += values[:1]
        ax.plot(angles, values, linewidth=2, linestyle='solid',
                label=model_name, color=colors[i % len(colors)])
        ax.fill(angles, values, alpha=0.1, color=colors[i % len(colors)])

    plt.legend(loc='upper right', bbox_to_anchor=(0.1, 0.1))
    plt.title(title, size=20, y=1.1)
    plt.tight_layout()
    plt.show()
radar_chart(Model_accuracy, 'Model Performance Comparison')
```



Digital Egypt Pioneers

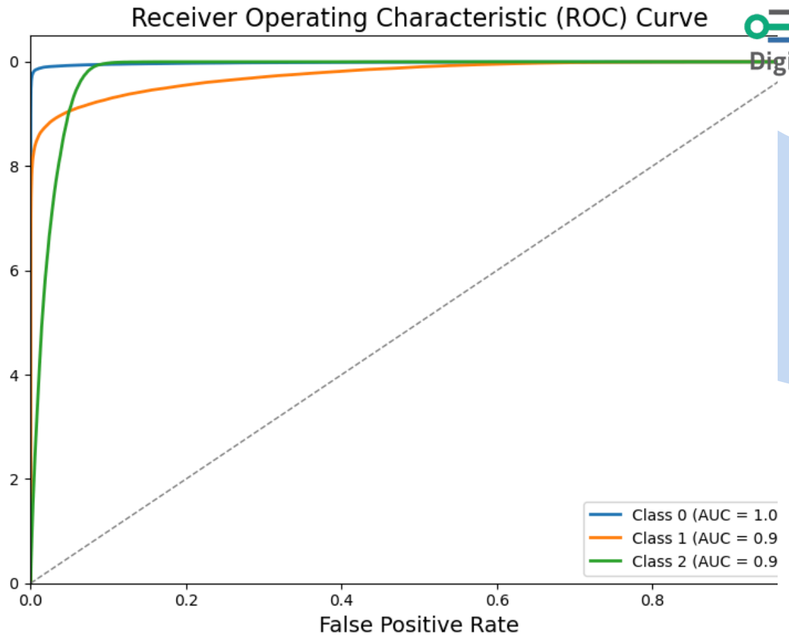
- A radar plot was used to compare the performance of the models.
- The evaluation includes four metrics: accuracy, precision, recall, and F1 score.
- The plot shows that Random Forest clearly outperforms the other models on all metrics,
- while Logistic Regression and GaussianNB performed similarly and significantly lower.

ANALYTICS ALCHEMISTS

Performance

```
from sklearn.metrics import classification_report

print(f"Classification Report for {best_model_name}:\n")
print(classification_report(y_test, y_pred))
```

```
Classification Report for RandomForestClassifier:

              precision    recall  f1-score   support

       Fatal       0.99      0.99      0.99     52655
     Serious       0.98      0.85      0.91     52655
      Slight       0.87      0.98      0.92     52655

    accuracy                           0.94    157965
   macro avg       0.95      0.94      0.94    157965
weighted avg       0.95      0.94      0.94    157965
```



Digital Egypt Pioneers

- The model achieved an overall accuracy of 94%.
- Excellent performance on fatal cases (precision and recall ≈ 99%).
- Good performance on serious cases, but low recall (85%), meaning the model misses some serious incidents.
- The model tends to fall into the light category (high recall 98%), but sometimes misclassifies errors as minor.

- The ROC curve illustrates the model's ability to distinguish between accident severity classes:
- Class 0 (Slight) achieved an AUC of 1.00, indicating perfect discrimination.
- Class 1 (Serious) and Class 2 (Fatal) both achieved AUC scores of 0.98, showing excellent classification performance.

ANALYTICS ALCHEMISTS

RECOMMENDATIONS

Digital Egypt Pioneers

- Based on the exploratory data analysis (EDA) of road traffic accidents, a number of practical recommendations were extracted that aim to reduce the number and severity of accidents by focusing on the most influential factors. The most important of these recommendations are:

1. Improving lighting and road conditions

- Installing strong lighting in high-risk areas, especially in areas where accidents occur during the dark.
- Improving road maintenance during adverse weather conditions (rain, fog, ice)

2. Improving intersection management

- Installing traffic signals at uncontrolled intersections.
- Using roundabouts at intersections with high accident rates.

ANALYTICS ALCHEMISTS

RECOMMENDATIONS

Digital Egypt Pioneers

3. Speed control

- Strictly enforcing speed limits, especially on highways and residential areas.
- Using speed cameras and speed bumps at critical locations

4. Monitoring Heavy and Private Vehicles

- Conducting rigorous periodic inspections of heavy vehicles.
- Educating heavy transport drivers about risks and maintaining legal loads.

5. Targeting high-accident times

- Increasing traffic control in the afternoon, especially on Fridays and Saturdays.
- Awareness campaigns on night driving and reducing dangerous driving The influence of alcohol or fatigue

ANALYTICS ALCHEMISTS

RECOMMENDATIONS

Digital Egypt Pioneers

6. Developing Predictive Data

- Building predictive models using artificial intelligence techniques to identify high-risk areas and times
- Using these models to allocate resources and improve proactive response.

7. Road Maintenance Infrastructure

- Periodic road maintenance, addressing slippery surfaces, potholes, and physical hazards.
- Providing clear warning signs for areas where roadwork is being carried out.

8. Focusing on Small Vehicles

- Requiring intensive training for new drivers.
- Promoting awareness about safety systems such as ABS and ESP.

ANALYTICS ALCHEMISTS

RECOMMENDATIONS

Digital Egypt Pioneers

9. Seasonal Analysis

- Intensify awareness and monitoring during specific months, such as March (increase) and December (decrease).

10. Continuous Evaluation

- Monitor the impact of recommendations through performance indicators (KPIs)
- Update and analyze data periodically to continuously adapt strategies.

Conclusion

Focusing efforts on (speed, lighting, intersections, heavy vehicles, and data analysis) can significantly contribute to reducing accidents. Data-driven decisions are key to improving traffic safety.

ANALYTICS ALCHEMISTS

Conclusion

Digital Egypt Pioneers

Analyzing road accident data is a fundamental step toward understanding the factors contributing to accident occurrence and severity. This analysis highlights the impact of speed, lighting conditions, intersection type, road conditions, and weather conditions in determining accident severity. The results demonstrate the importance of adopting proactive, data-driven strategies to enhance traffic safety.

Relying on intelligent data analysis not only helps reduce the number of accidents, but also helps guide traffic policies and allocate resources more efficiently, thus enhancing the protection of lives and achieving a safer traffic environment for all.

ANALYTICS ALCHEMISTS

Digital Egypt Pioneers

THANK YOU!

By / Analytics Alchemists