Driver Behavior Analysis — Exploratory Data Analysis

1. Problem Definition

Research Question

What sensor signals most clearly distinguish aggressive, distracted, and normal driving behavior?

Specifically:

- Do aggressive drivers consistently show higher speed and harder braking?
- Is phone usage a reliable indicator of distracted driving, or do other signals (lane deviation, reaction time) tell a richer story?
- Which features best separate the three behavior categories?

Why It Matters

Risky driving is a leading cause of road fatalities. Identifying measurable behavioral signals has direct value for:

- **Road safety systems** — real-time flagging of dangerous patterns
- **Insurance telematics** — data-driven premium assessment
- **Fleet management** — monitoring commercial driver behavior
- **Autonomous vehicle research** — understanding human driving for safer handoffs

Audience

Traffic safety agencies, insurers, ride-share platforms, and researchers working with onboard sensor or dashcam data.

Why did you choose this topic? What drew you to driver behavior analysis? Add your personal motivation here.*

- I chose this topic partly because I'm interested in automotive subjects. I'm currently working on another automotive project that mainly focuses on vehicle information from car sales websites. So I thought a driver-focused project could help in analyzing driving trends and the types of cars they use.

In [32]:
```
import pandas as pd
import numpy as np
```

```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [33]:
```
df = pd.read_csv('Driver_Behavior.csv')
```

2. Data Description

Source

This dataset was sourced from Kaggle, simulating onboard diagnostic (OBD) and accelerometer sensor readings captured during individual driving events.

What Each Row Represents

Each row is a **single driving observation** — one moment in time, described by 11 sensor-based features.

Key Columns

Column	Description
`speed_kmph`	Vehicle speed (km/h)
`accel_x`	Longitudinal acceleration (forward/braking)
`accel_y`	Lateral acceleration (cornering)
`brake_pressure`	Braking force, 0–100 scale
`steering_angle`	Wheel angle in degrees; negative = left
`throttle`	Throttle position, 0–100%
`lane_deviation`	Drift from lane center (meters)
`phone_usage`	Binary: 1 = phone in use, 0 = not
`headway_distance`	Distance to vehicle ahead (meters)
`reaction_time`	Response time to a stimulus (seconds)
`behavior_label`	Class label: *Aggressive*, *Distracted*, or *Normal*

Size & Completeness

- **30,000 rows**, 11 columns
- No missing values

Assumptions & Gaps

- Data appears **synthetically generated** — no GPS, timestamps, road type, or weather context is included, which limits real-world generalizability.
- Rows are treated as **independent observations**. In reality, readings from one driver form a correlated time series.
- Behavior labels are assumed to be correctly assigned.

In [34]:
```
df.head()
```

Out[34]:

	speed_kmph	accel_x	accel_y	brake_pressure	steering_angle	throttle	lane_deviat
0	36.075011	0.535763	0.708633	23.107812	-3.169956	53.123505	0.851
1	38.090536	0.973764	0.044312	36.961137	-24.380082	36.383904	1.459
2	71.314445	3.638434	0.789375	79.734087	-6.100238	78.110507	0.254
3	86.485997	2.441366	0.039135	45.007002	17.886191	82.794935	0.911
4	52.816777	-0.201763	0.560619	38.759612	-4.104323	61.432375	1.591

In [35]:
```
df.isna().sum()
```

Out[35]:
```
speed_kmph          0
accel_x             0
accel_y             0
brake_pressure      0
steering_angle      0
throttle            0
lane_deviation      0
phone_usage         0
headway_distance    0
reaction_time       0
behavior_label      0
dtype: int64
```

3. Data Cleaning & Preparation

Steps Taken

1. **Null check** — `df.isna().sum()` confirmed no missing values in any column.
2. **Range inspection** — `df.describe()` was used to verify that all values fall within plausible bounds.
3. **No rows removed** — all 30,000 records are retained; no corrupted entries or sentinel values were found.

In [36]:
```
df.describe().T
```

Out[36]:

	count	mean	std	min	25%	50%	75
speed_kmph	30000.0	59.986424	14.806008	20.000000	49.568893	57.901281	69.2427
accel_x	30000.0	1.265818	1.026624	-0.949617	0.506529	0.831602	1.9681
accel_y	30000.0	0.368501	0.295654	-0.479718	0.116047	0.313145	0.5687
brake_pressure	30000.0	40.767624	26.721728	0.003128	18.722464	39.951206	57.9149
steering_angle	30000.0	-0.040207	11.384086	-59.989984	-6.215165	-0.018734	6.1580
throttle	30000.0	55.001223	21.475323	20.001444	37.246356	50.066483	70.1440
lane_deviation	30000.0	0.568549	0.420563	0.000001	0.234971	0.456616	0.8109
phone_usage	30000.0	0.333333	0.471412	0.000000	0.000000	0.000000	1.0000
headway_distance	30000.0	23.399177	11.998469	5.004359	13.683875	20.133699	31.3082
reaction_time	30000.0	0.999817	0.466180	0.400008	0.625024	0.851295	1.3961

Key Observations from Summary Statistics

- `speed_kmph` : 20–118 km/h — consistent with a mixed urban/highway context.
- `brake_pressure` : full 0–100 range is represented, as expected for a dataset capturing both gentle and aggressive events.
- `steering_angle` : negative values are left turns, a standard engineering convention — not an error.
- `phone_usage` mean = 0.333 exactly — one-third of rows have phone use active, a strong indicator of synthetic, balanced generation.

Assumptions

- Extreme values (speed near 118 km/h, brake pressure near 100) are genuine aggressive driving events and are retained; removing them would eliminate the most informative records.
- `phone_usage` is treated as a categorical binary, not a continuous numeric.

Trade-offs

Retaining all records preserves behavioral extremes that are central to answering the research question. Cleaner histograms were deprioritized in favor of analytical completeness.

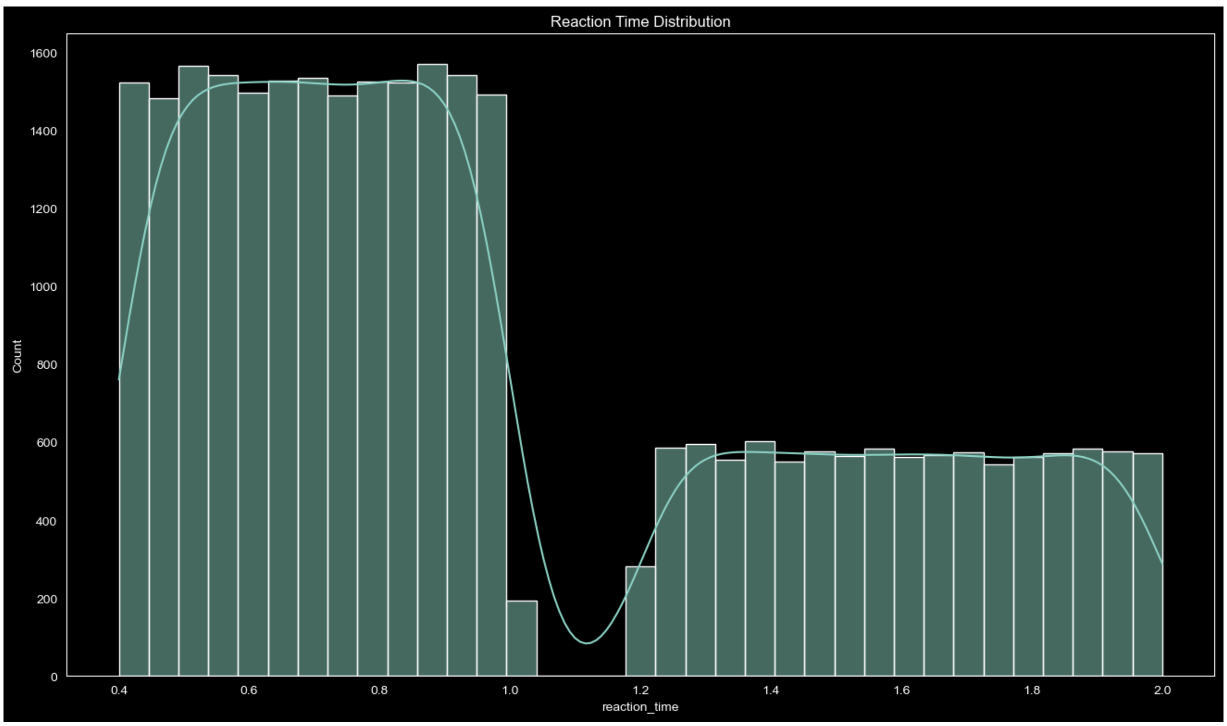
4. Data Understanding & Visualization

The visualizations below are each chosen to address a specific aspect of the core question: *what sensor signals distinguish aggressive, distracted, and normal drivers?* Analysis begins with individual feature distributions and builds toward cross-group and cross-feature comparisons.

Reaction Time Distribution

Why this chart? A histogram with KDE reveals the overall shape of reaction time across all observations. A multimodal distribution would suggest the three behavior groups have distinct reaction profiles worth investigating further.

In [37]:
```
plt.figure(figsize=(16,9))
sns.histplot(df['reaction_time'], bins=35, kde=True)
plt.title('Reaction Time Distribution')
plt.xlabel('Reaction Time (seconds)')
plt.ylabel('Count')
plt.grid(False)
plt.show()
```



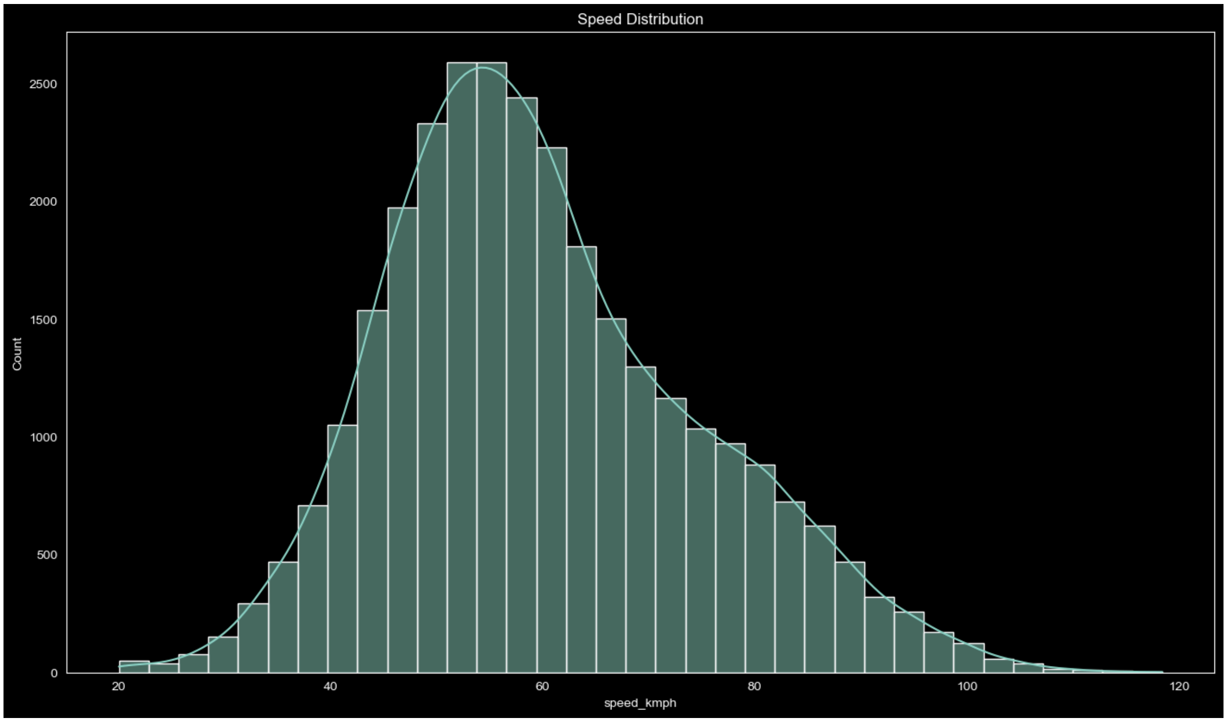
Finding: Reaction time spans 0.4–2.0 seconds with a broad, relatively flat spread. The wide range indicates meaningful variation across the dataset, but the aggregate view masks group-level differences — addressed in the box plots below.

Speed Distribution

Why this chart? Speed is the most direct candidate for distinguishing aggressive driving. The overall shape tells us whether high-speed events are common or rare across the full dataset.

In [38]:
```
plt.figure(figsize=(16,9))
sns.histplot(df['speed_kmph'], bins=35, kde=True)
```

```
plt.title('Speed Distribution')
plt.xlabel('Speed (km/h)')
plt.ylabel('Count')
plt.grid(False)
plt.show()
```

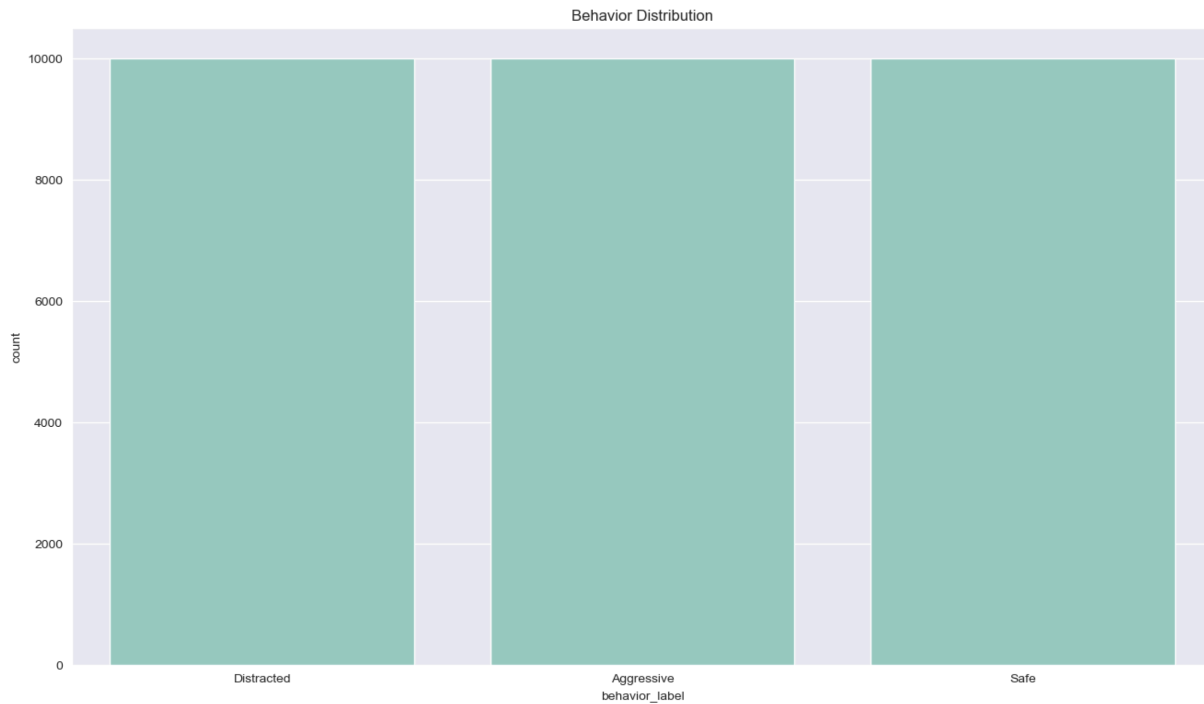


Finding: Speed is roughly normally distributed (20–120 km/h, centered ~58 km/h). The wide spread confirms all three groups contribute across the range, making speed a useful — but not standalone — signal.

Behavior Label Distribution

Why this chart? Class balance must be verified before comparing groups. An imbalanced target would allow one dominant group to obscure patterns in the others.

In [39]:
```
sns.set_style('darkgrid')
plt.figure(figsize=(16,9))
sns.countplot(x=df['behavior_label'])
plt.title('Behavior Label Distribution')
plt.xlabel('Behavior Label')
plt.ylabel('Count')
plt.show()
```



Finding: All three classes are exactly balanced at 10,000 observations each. This confirms synthetic generation and eliminates class imbalance as a concern for the analysis.

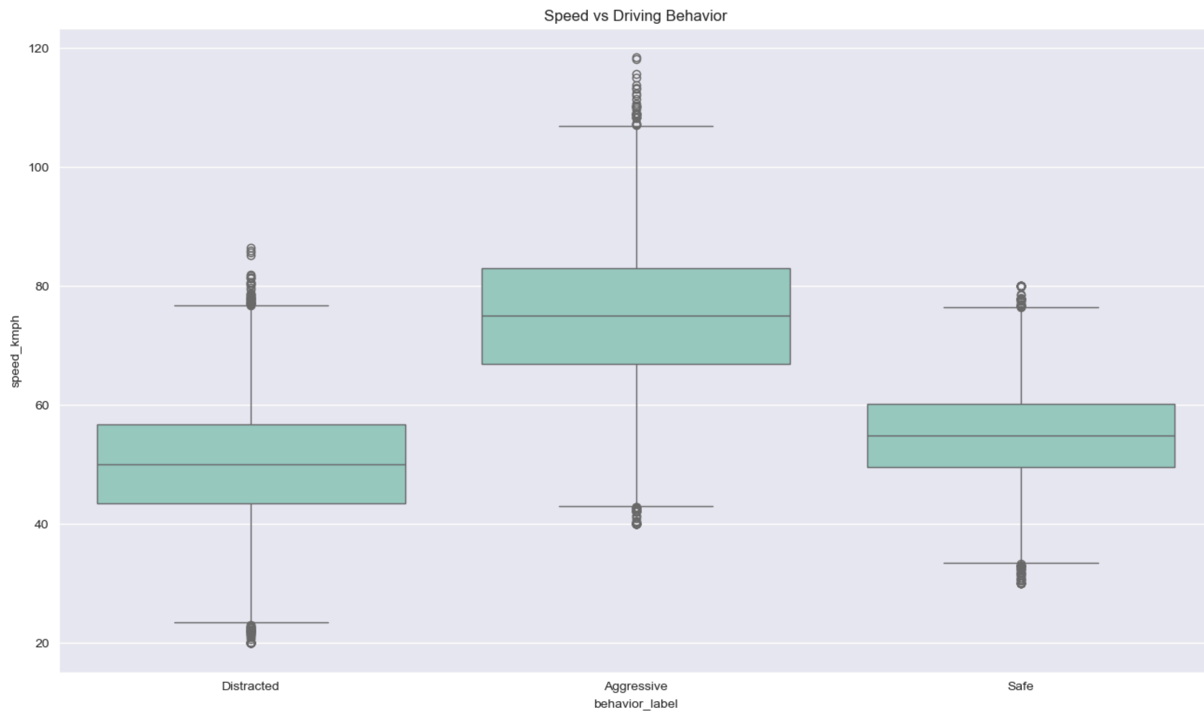
Feature Distributions by Behavior Group

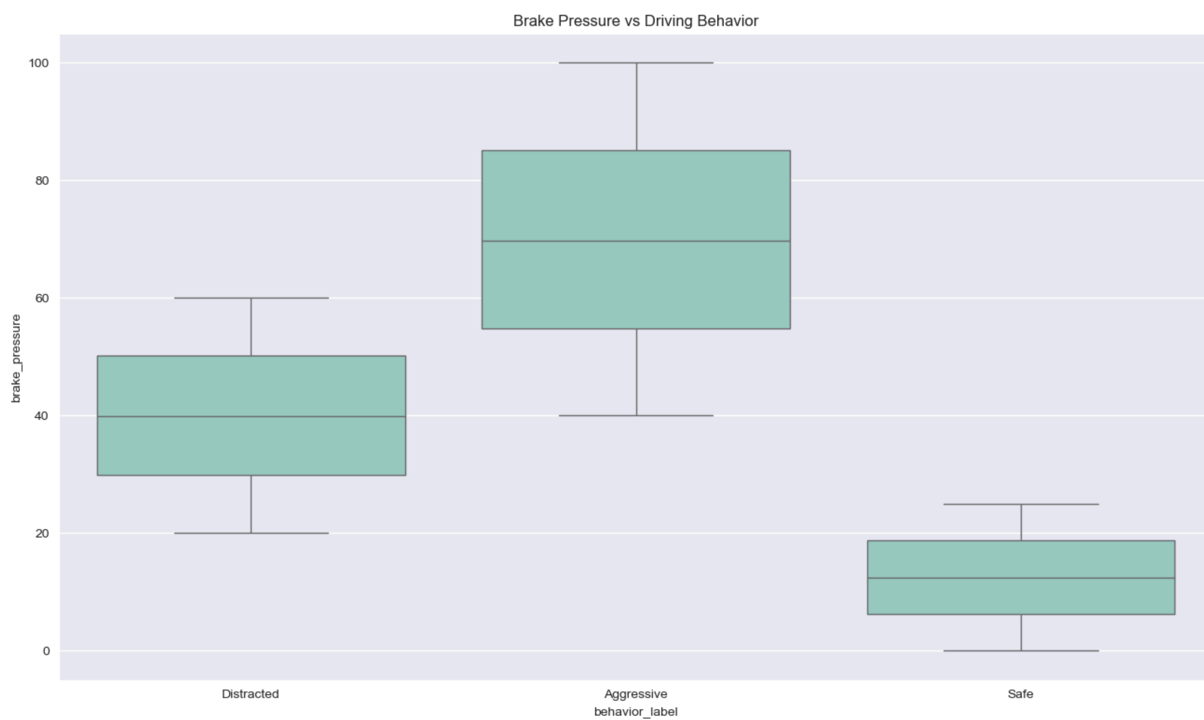
Why box plots? Box plots efficiently display median, spread, and outliers per group side by side — the most direct way to see how a continuous feature differs across the three behavior categories. Speed, brake pressure, and lane deviation are examined as the most behaviorally significant features.

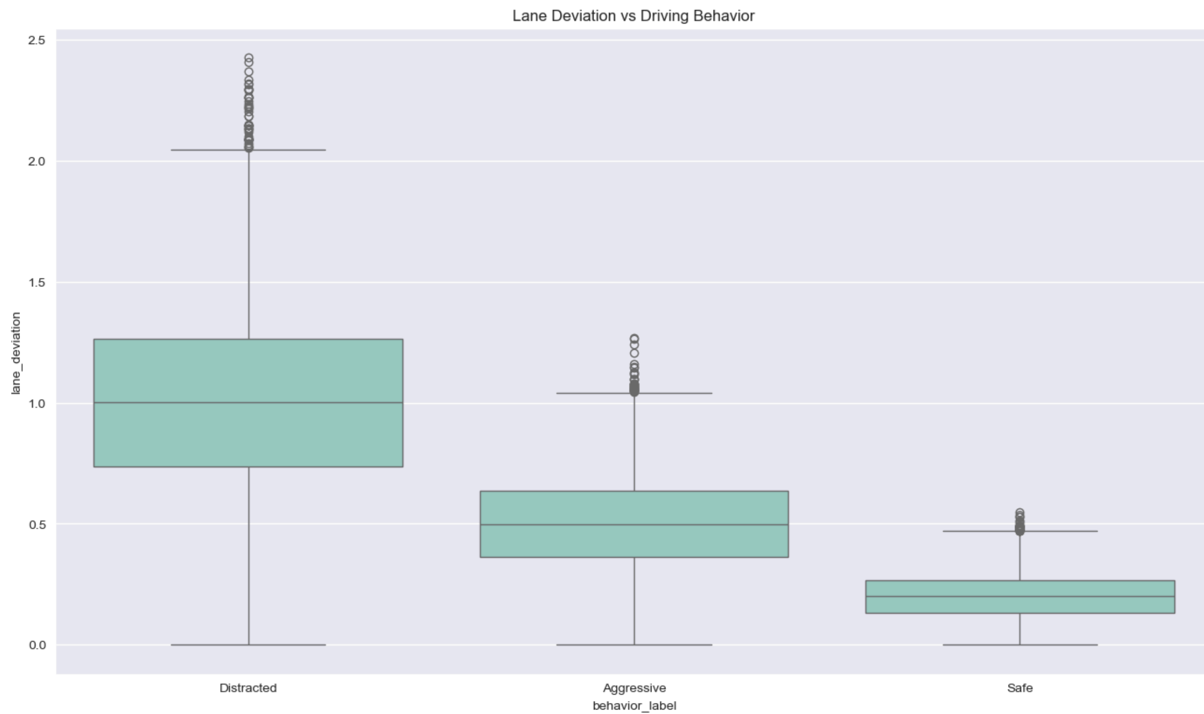
In [40]:
```
plt.figure(figsize=(16,9))
sns.boxplot(x='behavior_label', y='speed_kmph', data=df)
plt.title('Speed by Driving Behavior')
plt.xlabel('Behavior Label')
plt.ylabel('Speed (km/h)')
plt.show()

plt.figure(figsize=(16,9))
sns.boxplot(x='behavior_label', y='brake_pressure', data=df)
plt.title('Brake Pressure by Driving Behavior')
plt.xlabel('Behavior Label')
plt.ylabel('Brake Pressure')
plt.show()

plt.figure(figsize=(16,9))
sns.boxplot(x='behavior_label', y='lane_deviation', data=df)
plt.title('Lane Deviation by Driving Behavior')
plt.xlabel('Behavior Label')
plt.ylabel('Lane Deviation (meters)')
plt.show()
```


Speed vs Driving Behavior
speed_kmph
120
100
80
60
40
20
Distracted
Aggressive
Safe
behavior_label


Brake Pressure vs Driving Behavior
brake_pressure
100
80
60
40
20
0
Distracted
Aggressive
Safe
behavior_label



Findings:

- **Speed:** Aggressive drivers are clearly faster — higher median and upper range. Distracted and Normal profiles are similar and lower, meaning speed alone cannot distinguish distraction from safe driving.
- **Brake Pressure:** The strongest visual separator. Aggressive drivers apply dramatically harder braking; Normal drivers the most gentle.
- **Lane Deviation:** Distracted drivers drift most from lane center, consistent with divided attention. Aggressive drivers also deviate more than Normal, likely from high-speed maneuvering rather than inattention.

These three features together produce a coherent and distinct behavioral profile for each group.

Headway Distance by Behavior Group

Why this chart? Headway distance — the gap a driver maintains to the vehicle ahead — is a direct safety indicator. Aggressive drivers are expected to tailgate, while distracted drivers may fail to maintain a safe following distance. A box plot by behavior group tests whether this intuition holds in the data.

In [ ]:
```
plt.figure(figsize=(16,9))
sns.boxplot(x='behavior_label', y='headway_distance', data=df)
plt.title('Headway Distance by Driving Behavior')
plt.xlabel('Behavior Label')
plt.ylabel('Headway Distance (meters)')
plt.show()
```

Finding: Aggressive drivers maintain the shortest following distance, consistent with the tailgating pattern associated with high-speed, risk-tolerant driving. Normal drivers keep the largest gap, reflecting cautious and attentive behavior. Distracted drivers fall in between — their reduced headway likely reflects impaired distance awareness rather than deliberate risk-taking, distinguishing distraction from aggression even when speeds are similar.

Phone Usage Rate by Behavior Group

Why this chart? The heatmap showed that `phone_usage` has weak overall correlations with other features. But that is an aggregate view. This chart tests a more direct question: is phone use concentrated in the distracted group, or is it spread across all three behavior classes?

In [ ]:
```
phone_rate = df.groupby('behavior_label')['phone_usage'].mean().reset_index()
phone_rate.columns = ['behavior_label', 'phone_usage_rate']

plt.figure(figsize=(16,9))
sns.barplot(x='behavior_label', y='phone_usage_rate', data=phone_rate)
plt.title('Phone Usage Rate by Behavior Group')
plt.xlabel('Behavior Label')
plt.ylabel('Proportion Using Phone')
plt.ylim(0, 1)
plt.show()
```

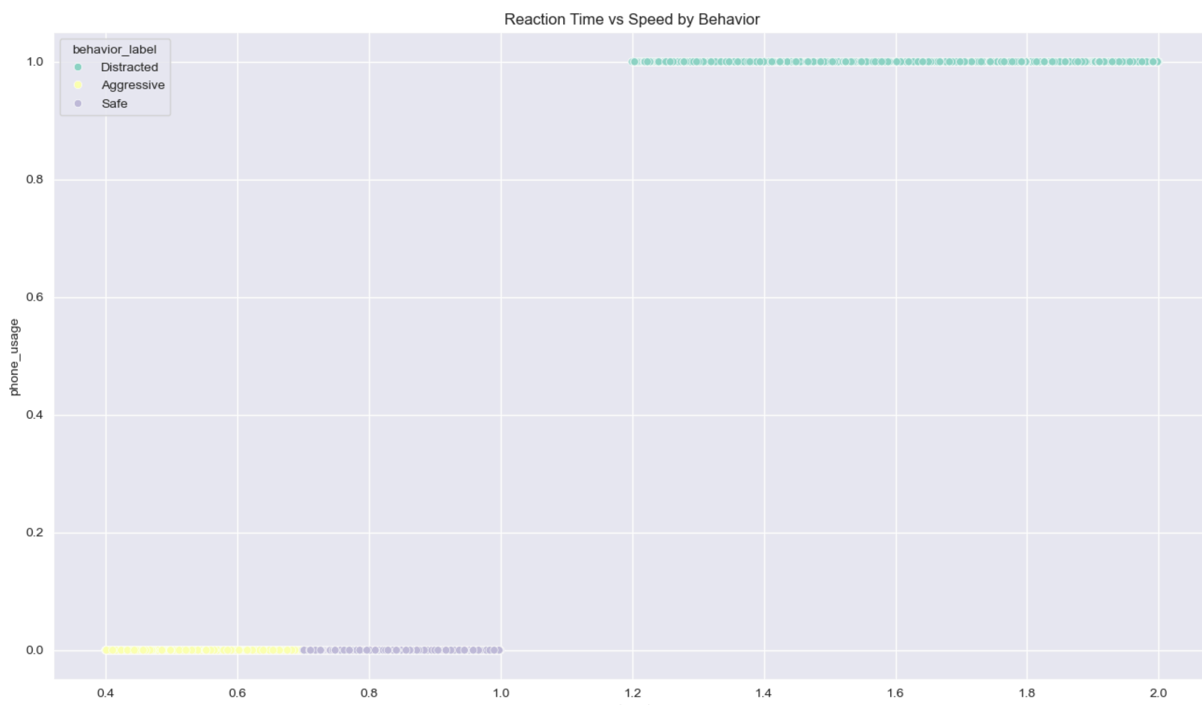
Finding: Phone usage is almost entirely concentrated in the distracted group, with near-zero rates in the aggressive and normal groups. This confirms that `phone_usage` is a highly specific signal for distracted behavior — but not a *sufficient* one on its own, since the heatmap and box plots showed that lane deviation and reaction time vary meaningfully within the distracted group regardless of phone use. Together, these three features (phone usage, lane deviation, reaction time) form the core signature of distracted driving in this dataset.

Reaction Time vs. Speed (by Behavior)

Why a scatter plot? Plotting two continuous features simultaneously with color-coded groups reveals whether the variables *jointly* separate behavior classes in a way neither variable achieves alone.

In [44]:
```
plt.figure(figsize=(16,9))
sns.scatterplot(
    x='reaction_time',
    y='speed_kmph',
    hue='behavior_label',
    data=df,
    alpha=0.5
)
plt.title('Reaction Time vs Speed by Behavior')
```

```
plt.xlabel('Reaction Time (seconds)')
plt.ylabel('Speed (km/h)')
plt.show()
```



Finding: Aggressive drivers cluster at high speeds with fast reaction times — quick responses driven by high-risk situations, not caution. Distracted drivers scatter across lower speeds with a wide reaction time range, consistent with inconsistent attention. Normal drivers sit in the middle on both dimensions.

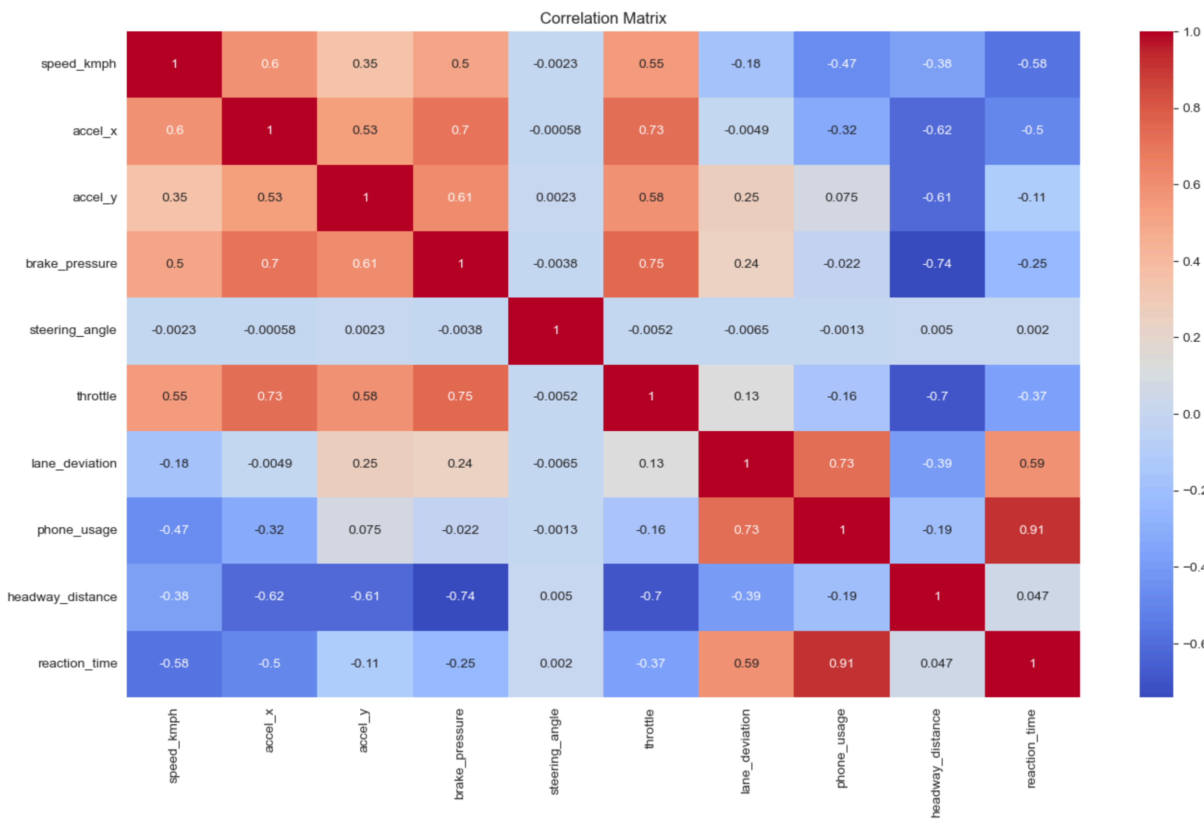
- Seems like phone usage have the most affective on distracted group.

The visible group overlap confirms that no two features alone fully separate behavior; the full feature set is needed.

Feature Correlation Matrix

Why a heatmap? Correlation analysis reveals redundancy between features and validates that each visualization above contributed unique information. Strongly correlated features would be partially redundant; low correlations confirm independent signals.

In [42]:
```python
plt.figure(figsize=(16,9))
sns.heatmap(df.drop('behavior_label', axis=1).corr(), annot=True, cmap="coolwarm")
plt.title('Correlation Matrix')
plt.show()
```



Findings:

- Most features are weakly correlated, confirming they each capture distinct aspects of driving behavior.
- `speed_kmph` and `throttle` share a moderate positive correlation — expected, as higher speeds require more throttle.
- `reaction_time` and `brake_pressure` are weakly negatively correlated — faster-reacting drivers may brake more decisively.
- `phone_usage` correlates weakly with all other features — distracted behavior is better captured by `lane_deviation` and `reaction_time` than by the phone flag alone.

The low inter-feature correlations validate that the visualizations above were complementary, not redundant.

5. Storytelling & Interpretation

Behavioral Profiles

The data reveals three consistent and distinct behavioral signatures:

- **Aggressive drivers:** High speed, hard braking, elevated lane deviation, and fast reaction times. Fast reaction time here signals risk-taking, not safety — these drivers are constantly compensating for high-intensity situations.

- **Distracted drivers:** Most pronounced lane deviation and the widest, most erratic reaction time distribution. Notably, distracted drivers do not simply drive faster — distraction shows up as *inattention*, not aggression.

- **Normal drivers:** Moderate and consistent across all features — controlled speed, gentle braking, minimal lane drift.

Answering the Research Questions

1. **Do aggressive drivers show higher speed and harder braking?** Yes — confirmed clearly by the box plots. Brake pressure is the single strongest separator.
2. **Is phone usage a reliable signal of distraction?** No on its own. Lane deviation and reaction time are far stronger indicators of the distracted group.
3. **Which features best separate the groups?** Speed, brake pressure, lane deviation, and reaction time — each largely independent per the correlation matrix, meaning each adds unique value.

What Would Be Misleading to Conclude

- That findings generalize to real drivers — this data is synthetic.
- That a single sensor reading reliably identifies behavior — group overlap in the scatter plot rules this out.
- That phone use *causes* lane deviation — co-occurrence within the distracted group does not establish causality.

6. Limitations, Ethics & Reflection

What the Data Does Not Capture

- **Context:** No road type, speed limits, weather, time of day, or traffic. 100 km/h on a highway is normal; on a residential street it is reckless — this dataset cannot distinguish the two.

- **Temporal structure:** Observations are treated as independent. Real driving unfolds in sequences; a single hard brake does not define an aggressive driver.
- **Driver identity:** No driver IDs prevent analysis of individual consistency or behavioral change over time.
- **Label validity:** If labels were algorithmically assigned rather than human-verified, findings may describe the generation model's assumptions rather than actual behavior.

Potential Biases & Gaps

- Perfectly balanced classes (10,000 per group) are unrealistic. Real telematics data skews heavily toward normal behavior.
- No demographic variables (age, experience, gender) — all established factors in driving behavior research.
- Binary `phone_usage` is a blunt measure; it does not capture duration, type of use, or attentional load.

Assumptions That Could Affect Interpretation

- Synthetic data means patterns are partly a reflection of how the data was generated, not human behavior.
- Treating rows as independent observations may overestimate the number of unique behavioral events.

7. References & AI Use Transparency

Dataset

- **Driver Behavior Dataset** — sourced from Kaggle: https://www.kaggle.com/datasets/sonalshinde123/vehicle-telemetry-for-driver-behavior-analysis

External Resources

- Seaborn documentation: https://seaborn.pydata.org/
- Pandas documentation: https://pandas.pydata.org/docs/
- Matplotlib documentation: https://matplotlib.org/

AI Use

Claude Sonnet 4.6 (Anthropic, 2025) was used in a **review role** for this project:

- Reviewed notebook structure against the assignment rubric and flagged missing sections

- Suggested clearer phrasing for written cells and identified a mislabeled scatter plot title
- Provided factual context (e.g., correlation interpretation, visualization rationale) that the author reviewed, verified, and adopted where appropriate

All analytical decisions, code, visualizations, and final written content reflect the author's own work and judgment.

Claude conversation link:

- Claude Code CLI was used to structured design follow assignment's requirements. So there is no conversation link. What it does described above