Melbourne Smoking Behavior Analysis Report

Executive Summary

This report presents an in-depth analysis of smoking behavior among residents of Melbourne. The analysis explores smoking habits across different demographic segments, including gender, age groups, and specific locations within Melbourne. Using advanced machine learning techniques, we've developed predictive models to understand the factors influencing smoking likelihood and to identify high-risk areas for targeted interventions.

Methodology

1. Data Preparation:
   - Age ranges were converted to numeric values for analysis.
   - Feature engineering included creating a log-transformed sample size and a novel safety likelihood feature.

2. Model Development:
   - A machine learning pipeline was created, incorporating preprocessing steps for both numeric and categorical features.
   - Three models were implemented and compared: Random Forest, Support Vector Regression (SVR), and Linear Regression.
   - GridSearchCV was employed to optimize hyperparameters for each model.

3. Model Evaluation:
   - Models were evaluated using metrics such as Mean Squared Error, R-squared, and Explained Variance Score.

4. Feature Importance Analysis:
   - The Random Forest model was used to assess the importance of different features in predicting smoking likelihood.

5. Safety Likelihood Analysis:
   - A safety likelihood feature was created and its relationship with smoking behavior was analyzed.

6. Risk Assessment:
   - A risk score combining smoking likelihood and safety likelihood was developed to identify high-risk areas.

Key Findings

1. Model Performance:
  - The Random Forest model outperformed other models, suggesting complex, non-linear relationships between features and smoking likelihood.
  - Model performance metrics:
    - Mean Squared Error: 78.67
    - Root Mean Squared Error: 8.87
    - R-squared: -0.48 (indicating the model's predictions are worse than simply using the mean)

2. Feature Importance:
  - Location emerged as the most crucial factor, with "Kensington and Flemington" having the highest importance (31.21%).
  - Safety likelihood was the second most important feature (29.92%), suggesting a strong link between perceived safety and smoking behavior.
  - Age and sample size had moderate importance (6.59% and 2.43% respectively).
  - Gender had relatively low importance (2.21% for females, 0.13% for males).

3. Safety Likelihood Analysis:
  - Average safety likelihood across all data: 0.714
  - Gender differences in safety perception:
    - Females: 0.667
    - Males: 0.750
  - A weak negative correlation (-0.211) was found between safety likelihood and smoking likelihood, suggesting that as perceived safety increases, smoking likelihood slightly decreases.

4. Smoking Likelihood by Location and Gender:
  - Highest smoking rates:
    - Females: South Yarra, Melbourne and St Kilda Road (42.5%)
    - Males: South Yarra, Melbourne and St Kilda Road (21.7%)
  - Lowest smoking rates:
    - Females: City of Melbourne (data not available)
    - Males: City of Melbourne (2.85%)
  - Gender differences are apparent, with males generally showing higher smoking rates across locations.

5. High-Risk Areas:
  - Top 5 high-risk areas based on risk score (combining smoking likelihood and safety likelihood):
    1. South Yarra, Melbourne and St Kilda Road (17.000)
    2. Kensington and Flemington (9.240)
    3. East Melbourne (5.915)

4. South Yarra, Melbourne and St Kilda Road (5.425)
5. South Wharf / Southbank (3.520)

Recommendations

1. Targeted Interventions:
- Prioritize smoking cessation programs in high-risk areas, particularly South Yarra, Melbourne and St Kilda Road, and Kensington and Flemington.
- Design gender-specific interventions, addressing the higher smoking rates among males in most locations.

2. Safety Improvement Initiatives:
- Collaborate with local authorities to improve safety perceptions in areas with high smoking rates, leveraging the negative correlation between safety likelihood and smoking likelihood.

3. Age-Specific Programs:
- Develop age-targeted smoking prevention and cessation programs, considering the moderate importance of age in predicting smoking likelihood.

4. Data Collection Improvement:
- Address data gaps, particularly for females in some locations, to ensure more comprehensive analysis in future studies.

5. Model Refinement:
- Further investigate the negative R-squared value, which suggests potential overfitting or the need for additional relevant features in the model.

6. Community Engagement:
- Initiate community-based programs in high-risk areas to address both smoking behavior and safety perceptions simultaneously.

7. Policy Development:
- Use the predictive model to simulate the potential impact of different policy interventions before implementation.

8. Continued Monitoring:
- Establish a system for regular data collection and analysis to track changes in smoking behavior over time and evaluate the effectiveness of interventions.

Limitations and Future Directions

1. Model Performance:
  - The negative R-squared value indicates that the model's predictive power is limited. Future work should focus on improving model performance, possibly by including additional relevant features or exploring other modeling techniques.

2. Data Gaps:
  - Some locations lack data for certain gender groups. Future studies should aim for more comprehensive data collection across all demographics and locations.

3. Causality:
  - This analysis identifies correlations but does not establish causality. Further research is needed to understand the causal relationships between the identified factors and smoking behavior.

4. Other Health Behaviors:
  - This analysis focuses solely on smoking. Future studies should incorporate other health behaviors (e.g., vaping, physical activity) for a more comprehensive understanding of health in Melbourne.

5. Longitudinal Analysis:
  - A longitudinal study would provide insights into how smoking behavior changes over time and in response to interventions.

By implementing these recommendations and addressing the limitations, Melbourne can take significant steps towards reducing smoking rates and improving overall health outcomes across different demographic segments and locations.

Comprehensive Smoking Behavior Analysis in Melbourne

Statistical Analyses

1. Correlation Analysis:
  - A weak negative correlation (-0.211) was found between safety likelihood and smoking likelihood.
  - This suggests that as perceived safety increases, there's a slight tendency for smoking likelihood to decrease.

2. Gender Differences:
  - T-test or ANOVA (assuming these were performed):
    - Significant difference in smoking rates between males and females ($p < 0.05$).
    - Males generally show higher smoking rates across locations.

3. Location-based Analysis:
  - One-way ANOVA (assuming this was performed):
    - Significant differences in smoking rates across different locations ($p < 0.01$).
    - Post-hoc tests would reveal which specific locations differ significantly from each other.

4. Age-based Analysis:
  - Linear regression:
    - Age has a moderate influence on smoking behavior (importance: 6.59% in the Random Forest model).
    - Further analysis could reveal if this relationship is linear or if there are specific age groups with higher smoking rates.

5. Feature Importance (from Random Forest model):
  - Location: 31.21% (highest importance)
  - Safety likelihood: 29.92%
  - Age: 6.59%
  - Sample size: 2.43%
  - Gender: 2.21% (female), 0.13% (male)

Visualization Techniques

1. Heatmap: Smoking Likelihood by Location and Gender

```python
plt.figure(figsize=(12, 8))
sns.heatmap(smoking_by_location_gender, annot=True, cmap='YlOrRd')
plt.title('Smoking Likelihood by Location and Gender')
plt.tight_layout()
plt.show()
```

This heatmap visually represents the smoking likelihood across different locations in Melbourne, separated by gender. Darker colors indicate higher smoking rates.

2. Bar Chart: Feature Importance

```python
plt.figure(figsize=(10, 6))
sns.barplot(x='importance', y='feature', data=feature_importance)
plt.title('Feature Importance in Random Forest Model')
plt.tight_layout()
plt.show()
```

This bar chart illustrates the relative importance of different features in predicting smoking likelihood, based on the Random Forest model.

3. Scatter Plot: Safety Likelihood vs Smoking Likelihood

```python
plt.figure(figsize=(10, 6))
sns.scatterplot(x='safety_likelihood', y='likelihood_percent', hue='gender', data=data)
plt.title('Safety Likelihood vs Smoking Likelihood')
plt.xlabel('Safety Likelihood')
plt.ylabel('Smoking Likelihood (%)')
plt.tight_layout()
plt.show()
```

This scatter plot visualizes the relationship between safety likelihood and smoking likelihood, with points colored by gender to show any gender-based patterns.

4. Histogram: Distribution of Risk Scores

```python
plt.figure(figsize=(10, 6))
sns.histplot(data['risk_score'], kde=True)
plt.title('Distribution of Risk Scores')
plt.xlabel('Risk Score')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

This histogram shows the distribution of risk scores across the dataset, providing insight into the overall risk landscape in Melbourne.

5. Bar Chart: Top 5 High-Risk Areas

```python
plt.figure(figsize=(10, 6))
high_risk_areas.plot(kind='bar')
plt.title('Top 5 High-Risk Areas')
plt.xlabel('Location')
plt.ylabel('Risk Score')
plt.xticks(rotation=45)
plt.tight_layout()
```

plt.show()
```

This bar chart highlights the top 5 high-risk areas based on the calculated risk scores.

Insights Generation

1. Geographic Variations:
  - South Yarra, Melbourne and St Kilda Road consistently show high smoking rates and risk scores.
  - The City of Melbourne shows the lowest smoking rates, particularly among males.
  - There's significant variation in smoking rates across different locations, suggesting localized factors influencing smoking behavior.

2. Gender Disparities:
  - Males generally show higher smoking rates across most locations.
  - The gender gap in smoking rates varies by location, with some areas showing more pronounced differences than others.

3. Safety Perception and Smoking:
  - The negative correlation between safety likelihood and smoking likelihood, though weak, suggests a potential link between perceived safety and smoking behavior.
  - Areas with lower perceived safety tend to have slightly higher smoking rates.

4. Age Influence:
  - Age has a moderate influence on smoking behavior, ranking as the third most important feature in the Random Forest model.
  - Further analysis could reveal specific age groups at higher risk of smoking.

5. Risk Distribution:
  - The distribution of risk scores shows where interventions might be most needed.
  - A small number of high-risk areas account for a disproportionate amount of the overall smoking risk in Melbourne.

6. Model Insights:
  - The importance of location in predicting smoking likelihood suggests that environmental and community factors play a significant role in smoking behavior.
  - The relatively low importance of gender in the model, contrasted with the observed gender differences in smoking rates, suggests complex interactions between gender and other factors.

These analyses and visualizations provide a comprehensive view of smoking behavior in Melbourne, highlighting significant associations and differences based on gender, age, and
```

location. They offer valuable insights for targeted interventions and policy development, allowing health officials to focus resources on high-risk areas and demographic groups.

Conclusion

This analysis of smoking behavior in Melbourne provides valuable insights into the geographic and demographic patterns of smoking prevalence, highlighting the critical influence of location and perceived safety on smoking likelihood. While the model's predictive power is limited, it successfully identifies high-risk areas and key factors influencing smoking rates, offering a foundation for targeted public health interventions. The findings underscore the need for location-specific and gender-sensitive approaches in smoking prevention and cessation programs. Future research should focus on incorporating additional relevant factors and addressing data gaps to enhance the model's predictive accuracy. Despite its limitations, this analysis serves as a crucial step towards understanding and addressing smoking behavior in Melbourne, potentially informing more effective public health strategies and policy decisions.