Exploratory Data Analysis (EDA) of the Oral Cancer Prediction Dataset  
  
This is the exploratory data analysis (EDA) report on the oral cancer prediction dataset defining number and category attributes about oral cancer risk factors like age, smoking, alcohol consumption, and cancer stage. The analysis was meant to perform data cleaning operations, in terms of outlier and null values treatments, relationships between variables, and extraction of opinions.   
  
2. Data Cleaning  
2.1. Glimpse Towards the Dataset  
- The dataset is loaded and analyzed in terms of its structure.  
- Huge repository of attributes pertaining to oral cancer is stored.  
  
2.2 Treatment of Missing Values   
- Some few columns also have missing values.   
- The missing values in the numerical attributes were replaced with median value.   
- For categorical attributes, missing values were furnished by using mode value.  
  
2.3. Removal of Duplicate Rows  
- Duplicate records were checked and removed, thus preserving the integrity of the data.   
  
2.4. Outlier Detection and Treatment  
- Outliers were detected using the interquartile range (IQR) method.  
- Data points beyond 1.5 times the IQR were wiped from the dataset.  
  
All categorical values are converted to lower case and stripped of whitespace.

3. Exploratory Data Analysis (EDA)

3.1. Univariate Analysis

-Summary Statistics

- Summary statistics for numerical variables of the data included mean, median, variance, and skewness.- Histograms displaying distributions of key numerical features were given.  
- Boxplots for the detection of outliers in important attributes were drawn.  
  
3.2. Bivariate Analysis  
- Correlation Analysis  
- A correlation matrix was generated to analyze correlations among numerical variables.  
- Strong relationships were found between smoking habits and risk factors.  
- Scatter Plots  
- Scatter plots were used to analyze trends of numerical attributes like age vs. cancer stage and alcohol consumption vs. cancer stage.  
- Boxplots  
- The distributions of numericals across categorical ones are compared using boxplots - a case in point will be different cancer stages.   
  
3.3. Feature endangers Testing  
- Scatter Plots  
- The relationships between several numerical variables were investigated with the aid of scatter plots.   
- Heatmaps  
- The heatmap representation of correlation needs to be sought for multiple variables.  
- Grouped Comparisons  
- Crossover effects on cancer risks were analyzed through grouping the categorical variables for smoking and alcohol use.

4. Trends  
- Age and stage: Getting older seems to be associated with advanced stages of cancer.  
Patterns and trends in alcohol use: When alcohol is combined with smoking, the chances of cancer develop are seen to increase noticeably.  
Outlier Removal Effect: The cleaning of outliers has had a favorable effect on the interpretability of statistical analysis.  
  
Conclusion  
The dataset is cleaned, and missing values managed to make possible reliable analysis. The exploratory analysis could provide insights into some of the critical risk factors leading to oral cancer. Prediction models and applications in machine learning can now be applied to this dataset after cleaning..