### **Group 3**

Junseo Choi, Sojeong Lee, Soyeon Park, Ho-Jae Song, Heewon Yang

**1. Upload file**

* csv: 1, docx: 2

### **2. Formulate a research question**

### **Theme:** Seasonal Effects on Crime Rates

1. **Research Question:** Do crime rates in Chicago show significant seasonal variation?
2. **Null Hypothesis (H0):** There is no significant seasonal variation in crime rates in Chicago.
3. **Alternative Hypothesis (H1):** There is significant seasonal variation in crime rates in Chicago.

**3. An outline of the slide presentation that details**

**(1) Work Flow**

1. **Introduction (Soyeon Park)**

* background information
* Why is it interesting/important?

1. **Dataset (Soyeon Park)**

* background information
* source

1. **EDA - (Junseo Choi,)**

* summary statistics
* visualization

1. **Hypothesis testing - (Junseo Choi, Sojeong Lee)**

* table of result
* interpretation
* visualization

1. **Modeling - (Heewon Yang, Hojae Song)**

* table of result
* interpretation
* visualization

1. **conclusion - (Hojae Song)**

* synthesis of results from all previous sections

1. **Implication (Sojeong Lee, Heewon Yang)**

* what the result and
* conclusion

1. what they mean
2. why they’re important
3. what they can be used for

**(2)Presentation details**

1. **Data preparation Module - Dataset (Soyeon)**

This module involves the process of wrangling data. It consists of 3 steps, loading the dataset and preprocessing by importing, tidying data to relabel and reorganize data, and transforming data by manipulating it to make it suitable for analysis**.**

1. **Exploratory data analysis (EDA) Module - EDA (Junseo)**

This module uses preprocessed data to summarize statistics and visualize the distribution; this will help discover patterns and outliers of the data and understand it through descriptive statistics. i.e. summarizing the season data by using mutate(), and using summarize and group\_by to find the mean, median, mode, and standard deviation. All this would help visualize and find data for the hypothesis test.

The reason why we should use EDA for this assignment is that EDA helps in summarizing the main data and finding patterns and relationships in the dataset. The first step to do EDA is to clean(summarize) the data to ensure the dataset is accurate so it can provide a reliable dataset for the future hypothesis test. Then summary statistics can be done to find the mean, median, mode and standard deviation that can help identify potential trends. The next step is to visualize the summarized dataset. Then create the linear model to find relationships within the data.

1. **Hypothesis testing Module - Hypothesis testing (Junseo, Sojeong)**

This module involves testing hypotheses related to the research questions and evaluating the relationships between variables through regression analysis. i.e. to answer the question if crime is related to seasons, creating a null hypothesis that states that crime does not relate to seasons and an Alternative hypothesis of crime is related to seasons. Because we are measuring the significant variation in crime rates across seasons without knowing which season might have higher or lower rates we should use a two tailed test

Null Hypothesis: Crime rates do not vary by season / Alternative Hypothesis: Crime rates vary by season.

Function: lm(incidents ~ Season, data = crime\_data\_selected)

Testing assumptions: Linearity, Independence, Homoscedasticity, Normality.

1. **Modeling Module - Modeling(Heewon, Hojae)**

(1)Regression model

This module aims to estimate how the crime rate changes as the season and other independent variables change. Examine the model following 3 linear model assumptions.

Function: lm(response variable ~ explanatory variable, data), tidy(), glance()

Testing :

1. Linearity: Observed vs Predicted plot: add\_predictions(), add\_residuals()
2. Nearly normal residuals: QQ plot
3. Constant variability: Residuals vs predicted plot

(2) Understanding and Prediction model

This module aims to make accurate predictions about future seasonal crime rate. We can use various machine learning algorithms like Generalized Linear Model (GLM) to build predictive models and evaluate their performance.

Function: glm(crime\_rate (dependent variable) ~ Season + Other\_Variables ( independent variable) , family = "poisson", data = crime\_data\_selected)