The Effect of Heavy Wind on Power Outage Duration

Overview

This project utilizes hypothesis testing to evaluate the characteristics of prolonged power outage. We used publicly avaliable major power outage data in the continental U.S. from January 2000 to July 2016 to perform statistical analysis. In this project, a major power outage is defined as a power outage that impacted at least 50,000 customers or caused an unplanned firm load loss of atleast 300MW. Our result suggests that heavy wind leads to longer durations of power outages compared to other categories of cause.

Introduction

Power outage is a severe yet common condition that we all have experienced in our life. When a power outage occurs, it could disrupt communications, water and transportation; services as grocery stores, gas stations, ATMs, banks, retail businesses would be paralyzed as well. And when a power outage is prolonged, it could even cause food spoilage and water contamination. Meanwhile, with rising global average temperature and increased emission of greenhouse gas, extreme weather events become more frequent and intense. Would more frequent extreme weather events result in more occurences of power outages? Would specific severe weathers, such as storms, leads to longer duration of outages?

The data used in this project can be found at Major Power Outage Risks. This dataset contains major outages, information on geographical location of the outages, regional climatic information, land-use characteristics, electricity consumption patterns and economic characteristics of the states affected by the outages.

Cleaning and EDA

In [2]:
```
import matplotlib.pyplot as plt
from scipy.interpolate import make_interp_spline, BSpline
import geopandas as gpd
import folium
from folium.features import GeoJsonTooltip
import numpy as np
import os
import pandas as pd
import seaborn as sns
import datetime as dt
%matplotlib inline
%config InlineBackend.figure_format = 'retina'  #Higher resolution figures
import warnings
warnings.filterwarnings('ignore') #suppress warnings
```

In [3]:
```
df = pd.read_excel('outage.xlsx', header = 5).iloc[1:,2:]
df['OUTAGE.DURATION'] = df['OUTAGE.DURATION'].astype(float)
df['OUTAGE.START'] = pd.to_datetime(df['OUTAGE.START.DATE']) + pd.to_timedelt
df['OUTAGE.RESTORATION'] = pd.to_datetime(df['OUTAGE.RESTORATION.DATE']) + pd
df.drop(columns = ['OUTAGE.START.DATE', 'OUTAGE.START.TIME',
                   'OUTAGE.RESTORATION.DATE', 'OUTAGE.RESTORATION.TIME'], inp
df = df[(df['CUSTOMERS.AFFECTED'] >= 50000) | (df['DEMAND.LOSS.MW'] >= 300)]
df.head()
```

Out[3]:

	YEAR	MONTH	U.S._STATE	POSTAL.CODE	NERC.REGION	CLIMATE.REGION	ANOMALY.LE
1	2011.0	7.0	Minnesota	MN	MRO	East North Central	-
3	2010.0	10.0	Minnesota	MN	MRO	East North Central	
4	2012.0	6.0	Minnesota	MN	MRO	East North Central	
5	2015.0	7.0	Minnesota	MN	MRO	East North Central	
6	2010.0	11.0	Minnesota	MN	MRO	East North Central	

5 rows × 53 columns

Geographic Correlation of Categories of Causes

From the dataset, we noticed that there are only 7 categories of all the events causing the major power outages. we want to investiagte what kind of role does cause play in our dataset? Specifically, is there a category that causes the majority of power outages?

In [4]:
```
df.groupby('CAUSE.CATEGORY').count().iloc[:,0].sort_values(ascending = False)
```

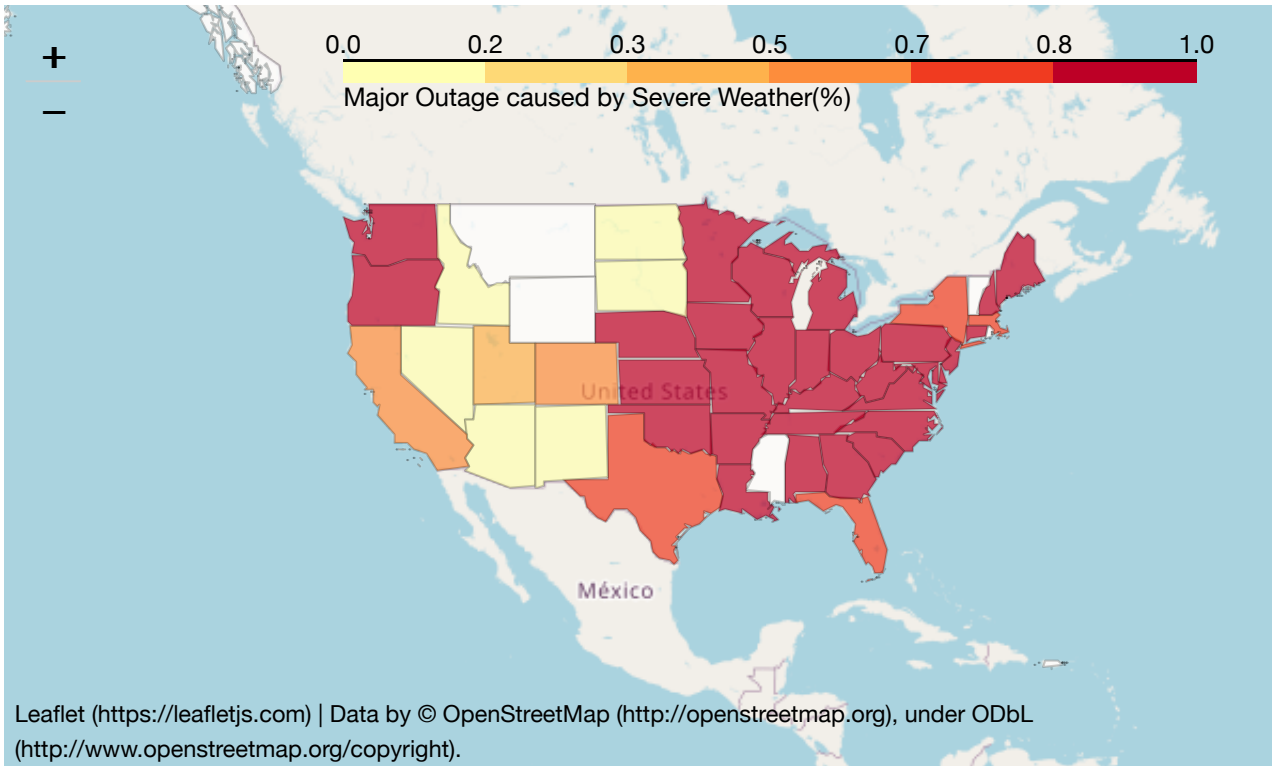
Out[4]:
```
CAUSE.CATEGORY
severe weather                    685
system operability disruption      76
equipment failure                  26
fuel supply emergency              16
public appeal                       7
intentional attack                  5
islanding                           2
Name: YEAR, dtype: int64
```

In the major power outages, severe weather events was the main cause of outages with a percentage of 83.8%. However, does all the states equally likely to experience events of severe weather? Or is there a regional factor correlates with the cause of outages? To further our study, we visualized this geospatial data using the frequency of severe weather events in the states.

In [5]:
```
#extract observations with severe weather events
temp = df[df['CAUSE.CATEGORY'] == 'severe weather'].groupby('U.S._STATE').cou
temp = ((temp.iloc[:, 0:1]/df.groupby('U.S._STATE').count().iloc[:,0:1]).fill
state_geo = gpd.read_file('geo.json')  #reads the geoJSON file
temp = temp.merge(state_geo, left_on = 'U.S._STATE', right_on = 'NAME')

#geospatial visualization
m = folium.Map(location=[40, -98], zoom_start = 4, title = 'Total Number of M
folium.Choropleth(
    geo_data = state_geo,
    name = "choropleth",
    data = temp,
    columns = ['U.S._STATE', 'YEAR'],
    key_on = 'properties.NAME',
    nan_fill_color = "White", # Use white color if there is no data for the s
    fill_color = 'YlOrRd',
    fill_opacity = 0.7,
    line_opacity = 0.2,
    highlight = True,
    legend_name = 'Major Outage caused by Severe Weather(%)',
).add_to(m)
m
```

Out[5]:



As observed in the chorepleth above, most U.S. states in the continent are experiencing nearly all power outages induced by severe weather, while states in South, West and Southweat climate regions have moderate to low percentage of outage caused by severe weather.

Correlation between Specific Severe Weather Types and Outage Durations

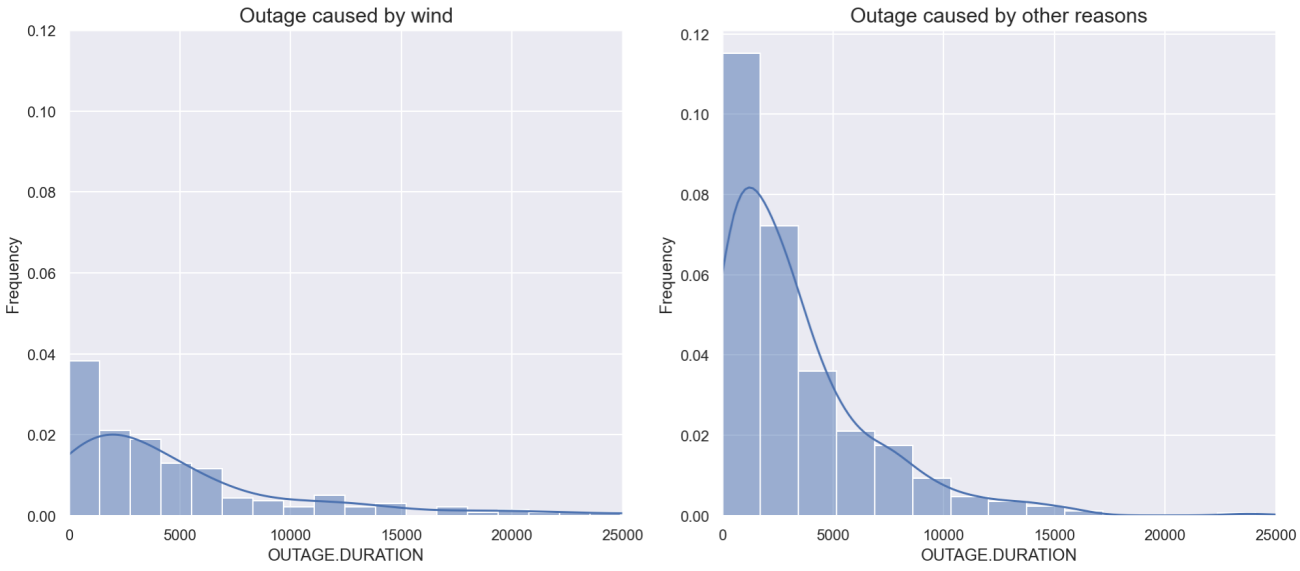
To investigate further into severe weather events, we are interested in the specific types of the events as well as their associated outage durations. As we expected, there are many events related to wind. We thus classified the specific causes of severe weather events into two main categories: related to heavy wind and unrelated to heavy wind. Our classification of heavy wind is due to the Beaufort Scale. Specifically, we included *heavy wind, storm, wind storm, hurricanes and tornadoes* as heavy wind events and the rest as unrelated to heavy wind.

To find the correlation between heavy wind events and outage durations, we started by visualizing this univariate data.

In [6]:

```
wind = ['heavy wind', 'storm', 'wind storm', 'hurricanes', 'tornadoes']
weather_df = df[df['CAUSE.CATEGORY'] == 'severe weather'] #extract severe wea
weather_df['WIND'] = weather_df['CAUSE.CATEGORY.DETAIL'].isin(wind)
weather_df.dropna(subset = ['CAUSE.CATEGORY']) #drop null values

#visualization
sns.set(rc={'figure.figsize':(13.5,6)})
fig, (ax1, ax2) = plt.subplots(1, 2)
sns.histplot(x = "OUTAGE.DURATION",   #histogram of outage duration with heav
             data = weather_df[weather_df['WIND']],
             kde = True,
             ax = ax1,
             stat = 'frequency',
             bins = 20)
sns.histplot(x = "OUTAGE.DURATION", #histogram of outage duration with non he
             data = weather_df[weather_df['WIND'] == False],
             kde = True,
             ax = ax2,
             stat = 'frequency',
             bins = 20)
ax1.set_title('Outage caused by wind', fontsize = 15)
ax2.set_title('Outage caused by other reasons', fontsize = 15)
ax2.set_xlim(0, 25000)
ax1.set_ylim(0, 0.12)
ax1.set_xlim(0, 25000)
fig.tight_layout()
```



Comparing the distribution of outage durations associated with and without heavy wind, we find out that graph on the left(heavy wind) is bimodal with one peak around 0-5000mins and another smaller peak around 10000-15000mins, whearas the graph on the right(non heavy wind) is unimodal with observations concentrated around 0-5000 mins. Since the two distribution aren't similar, we propose that there is a correlation between the specific cause detail of the outage and the duration of the outage.

The statistcis of the two distributions are:

In [7]:
```
#computes the averages of outage durations
t = weather_df.groupby('WIND')['OUTAGE.DURATION'].agg(['mean', 'median', 'cou
t.index = ['Non-Wind Induced Outage Duration', 'Wind Induced Outage Duration'
t
```

Out[7]:

	mean	median	count
Non-Wind Induced Outage Duration	3346.211066	2502.5	488
Wind Induced Outage Duration	4976.546961	3264.0	181

We noticed that wind induced outages have a higher mean and median duration compared to other outages, to verify if this has happened due to chance alone, we would perform a hypothesis test later.

Identify Missingness of Detailed Causes of Events

Before performing a hypothesis test, it is important to identify the type of missingness of the data that we are interested in to prevent bias in our conclusion.

Proportion of data missing in outage duration:

In [8]:
```
df['OUTAGE.DURATION'].isna().sum()/df.shape[0]
```

Out[8]:
```
0.023255813953488372
```

In [9]:
```
df[df['OUTAGE.DURATION'].isna()].groupby('CAUSE.CATEGORY.DETAIL').count().ilo
```

Out[9]:
```
CAUSE.CATEGORY.DETAIL
Coal              0
heavy wind        1
hurricanes        1
line fault        1
thunderstorm      0
wildfire          1
wind              1
winter storm      0
Name: MONTH, dtype: int64
```

The above series denotes the occurence of types of detailed causes when duration is missing, since only 2% of duration data is missing and it does not appears to depend on causes, we conclude that the missingness of outage duration is trivial and would ignore the null values when performing hypothesis testing.

Proportion of data missing in detailed causes of events:

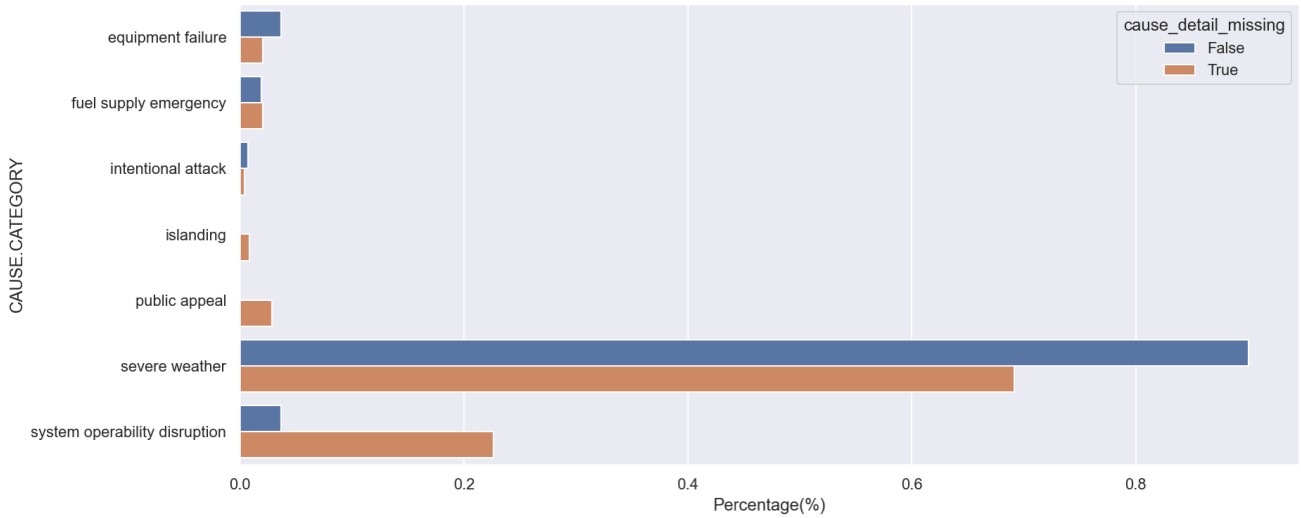
In [10]:
```
df['CAUSE.CATEGORY.DETAIL'].isna().sum()/df.shape[0]
```

Out[10]:
```
0.2974296205630355
```

In this dataset, the detail causes of events column expands on the general causes of event column, we thus analyze if the missingness in the column that we are interested in depend on the `CAUSE.CATEGORY` column first.

In [11]:
```
#calculates the percentage of missing and non-missing data in each category
assert_missingness = df.copy()
assert_missingness['cause_detail_missing'] = assert_missingness['CAUSE.CATEGO
emp_distributions = (
    assert_missingness
    .pivot_table(index='CAUSE.CATEGORY', columns='cause_detail_missing', valu
    .fillna(0)
    .apply(lambda x: x / x.sum())
)
emp_distribution_graph = emp_distributions.unstack().reset_index()

#visualization
ax = sns.barplot(data = emp_distribution_graph,
                 x = 0,
                 y = 'CAUSE.CATEGORY',
                 hue = 'cause_detail_missing')
ax.set_xlabel('Percentage(%)')
plt.show()
```



As shown in the horizontal bar graph above, there is a significant difference between the percentage of data missing when the cause category of an outage is "system operability disruption".

In order to test whether the missingness in the column `CAUSE.CATEGORY.DETAIL` depends on column `CAUSE.CATEGORY`, we decide to perform a permutation test on these two cloumns.

Null Hypothesis: the missingness in column `CAUSE.CATEGORY.DETAIL` does NOT depend on column `CAUSE.CATEGORY`
Alternative Hypothesis: the missingness in column `CAUSE.CATEGORY.DETAIL` depend on column `CAUSE.CATEGORY`

In [12]:
```
# calculate the observed tvd
observed_tvd = emp_distributions.diff(axis=1).iloc[:, -1].abs().sum() / 2
observed_tvd
```

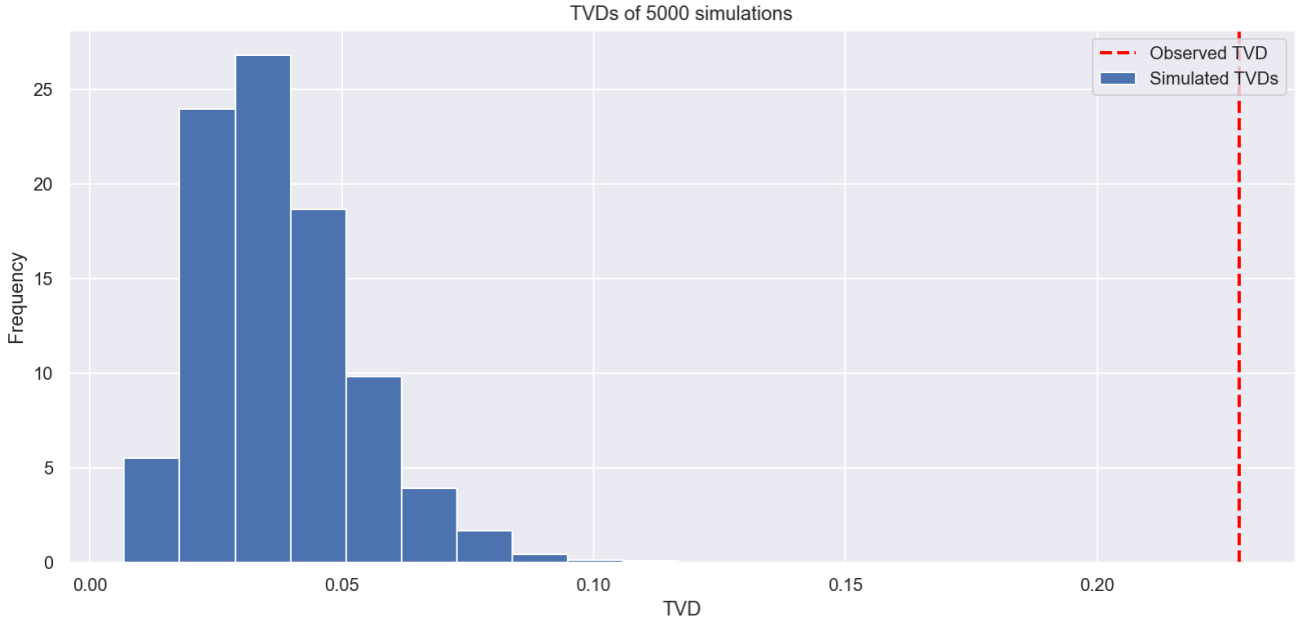
Out[12]: 0.22820148836409002

In [13]:
```
#perform a permutation test
shuffled = assert_missingness.copy()[['CAUSE.CATEGORY', 'cause_detail_missing

n_repetitions = 5000
tvds = []

for _ in range(n_repetitions):
    #Shuffling the column and assigning it to the DataFrame
    shuffled['CAUSE.CATEGORY'] = np.random.permutation(shuffled['CAUSE.CATEGO
    #Computing and storing the TVD
    pivoted = (
        shuffled
        .pivot_table(index='CAUSE.CATEGORY', columns='cause_detail_missing', 
        .fillna(0)
        .apply(lambda x: x / x.sum())
    )
    tvd = pivoted.diff(axis=1).iloc[:, -1].abs().sum() / 2
    tvds.append(tvd)
```

In [14]:
```
#visualization of 5000 simulations
ax = pd.Series(tvds).plot(kind = 'hist',
                      density = True,
                      ec = 'w',
                      bins = 10,
                      title = '',
                      label = 'Simulated TVDs')
plt.axvline(x = observed_tvd,
            color = 'red',
            linewidth = 2,
            label = 'Observed TVD',
            linestyle = '--')
plt.legend(loc = 'upper right')
plt.title('TVDs of 5000 simulations')
ax.set_xlabel('TVD')
plt.show()
```



P-value of this permutation test:

In [15]:
```
pval = np.mean(tvds >= observed_tvd)
pval
```

Out[15]: 0.0

p-value of 0.0 indicates that we reject our null hypothesis and claim that the missingness in column `CAUSE.CATEGORY.DETAIL` depend on column `CAUSE.CATEGORY`. This result is also illustrated in the histogram above as the TVDs of our simulated data are all on the left side of our observed TVD, which is indicated as the red line.

Next, in order to find out whether the missingness of `CAUSE.CATEGORY.DETAIL` depends on other columns, we decide to perform permutation tests on other columns to find what else contribute to the missingness of our target column. We decide to focus on columns of `ANOMALY.LEVEL`, `TOTAL.PRICE`, and `TOTAL.SALES` since the data for these columns are different for different outages.

Different from the `CAUSE.CATEGORY` column, these new columns have numeric data such that it is unreasonable to use TVD as their test statisitcs. We thus decided to use the absolute mean difference as the test statistic of their permutation test.

In [16]:
```
#get the other columns of interest
other_columns = ['ANOMALY.LEVEL', 'TOTAL.PRICE', 'TOTAL.SALES']
other_columns
```

Out[16]: ['ANOMALY.LEVEL', 'TOTAL.PRICE', 'TOTAL.SALES']

In [17]:
```
# define a function called permutation_miss which takes in the column to perm
# Returns a list to different in means
def permutation_miss(column, df):
    shuffled = df.copy()[[column, 'cause_detail_missing']]
    n_repetitions = 5000
    diff_means = []

    for _ in range(n_repetitions):
        # Shuffling the data and assigning it back to the DataFrame
        shuffled[column] = np.random.permutation(shuffled[column])
        # Computing and storing the absolute difference in means
        diff_mean = shuffled.groupby('cause_detail_missing')[column].mean().d
        diff_means.append(diff_mean)

    return diff_means
```

In [18]:
```
# perform permutation tests on all the column mentioned above.
# use difference in mean as test statistic
p_values = {}
diff_m = {}
for column in other_columns:
    # for each column, first find the observed test statistic than perform pe
    obs = assert_missingness.groupby('cause_detail_missing')[column].mean().d
    diff_means = permutation_miss(column, assert_missingness)
    p_values[column] = np.mean(diff_means >= obs)
    diff_m[column] = diff_means
p_values
```

Out[18]:
```
{'ANOMALY.LEVEL': 0.745, 'TOTAL.PRICE': 0.0046, 'TOTAL.SALES': 0.0008}
```

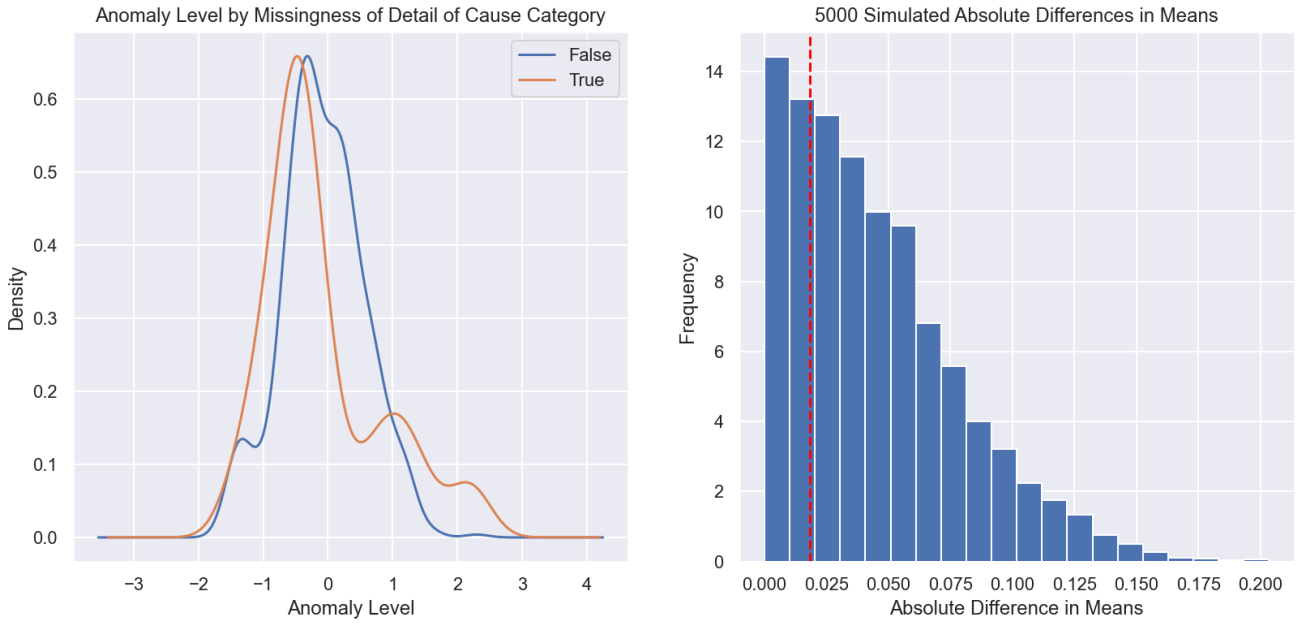
As shown in the output above, the p-values for column `TOTAL.PRICE` and `TOTAL.SALES` are 0.0046 and 0.0008 respectively and are all less than the statistically significant threshold of p = 0.05, indicating that the mssingness in our target data `CAUSE.CATEGORY.DETAIL` also depends on columns of total price and total sales.

On the other hand, the p-value for column `ANOMALY.LEVEL` is 0.745, which is greater then 0.05 threshold. We thus fail to reject our null hypothesis and claim that the missinness in our target data `CAUSE.CATEGORY.DETAIL` does NOT depend on column anomaly level.

We further visualized the correlation between anomaly level and the missingness of detailed cause of category below.

In [19]:

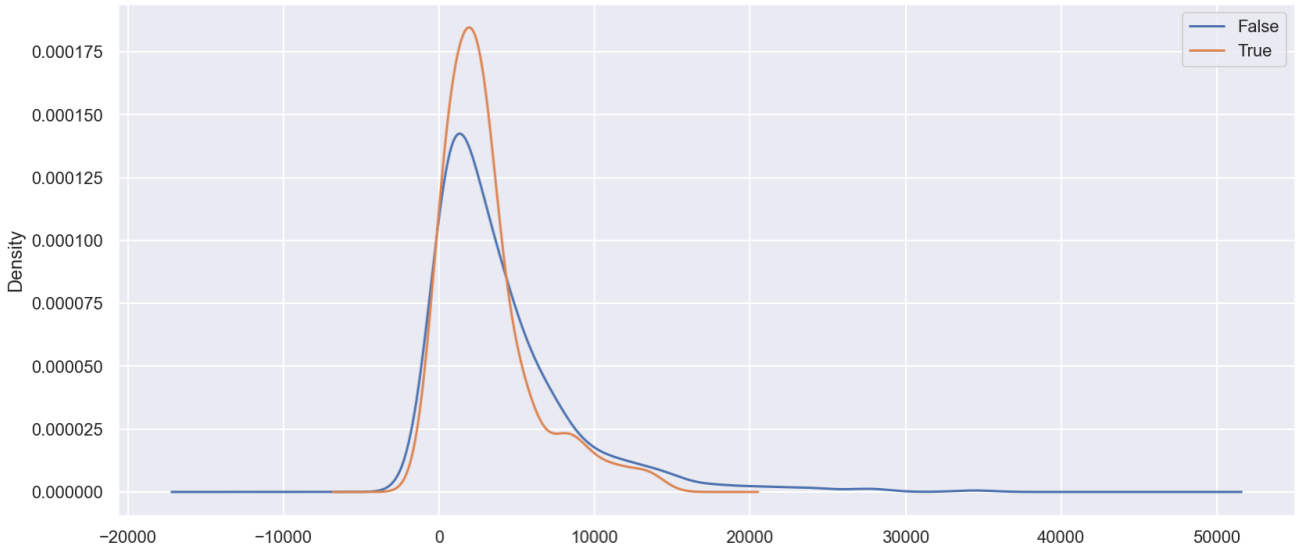
```
#plot out the two distributions
fig, (ax1, ax2) = plt.subplots(1, 2)
assert_missingness.groupby('cause_detail_missing')['ANOMALY.LEVEL'].plot(kind
                                                                         lege
                                                                         titl
                                                                         ax =
ax1.set_xlabel('Anomaly Level')
obs = assert_missingness.groupby('cause_detail_missing')['ANOMALY.LEVEL'].mea
ax2 = pd.Series(diff_m['ANOMALY.LEVEL']).plot(kind='hist',
                                              density=True,
                                              ec='w',
                                              bins=20,
                                              title='5000 Simulated Absolute
                                              ax = ax2)
ax2.set_xlabel('Absolute Difference in Means')
plt.axvline(obs,
            color='red',
            label='Observed Difference in Means',
            linestyle = '--')
plt.show()
```



The above two graphs illustrate the observed distributions of anomaly level when cause is missing/not missing(left), and the simulated distribution of absolute different in means(right). The left graph demonstrates that the similarity of the two observed distributions and the right graph further proves that missingness of detailed causes does not depend on anomaly level as majority of the simulated statistics goes beyond the observed statistic (portrayed in red line).

In general, we found that the missingness in column `CAUSE.CATEGORY.DETAIL` depends on the columns of `CAUSE.CATEGORY`, `TOTAL.PRICE` and `TOTAL.SALES` yet does not depend on columns `ANOMALY.LEVEL`. Overall, we conclude that the missingness of detailed cause category data is **MAR**. To decide how to handle the missing values while performing a hypothesis test against duration, we visualized the durations of outages when category detail is null and is not null.

In [20]:
```
weather_df['detail_missing'] = weather_df['CAUSE.CATEGORY.DETAIL'].isna()
weather_df.groupby('detail_missing')['OUTAGE.DURATION'].plot(kind='kde', lege
plt.show()
```



The visualization shows that the two distributions of outage durations are similar, that is, the missingness of detail cause category does not dependent on outage durations and we thus retain the null values in our hypotehsis testing.

Hypothesis Testing on Outage Duration and Cause of the Event

Null Hypothesis: In major outages caused by severe weather events, there is no association between the specific cause detail and outage duration

-- the high average duration of wind induced outages is due to chance alone.

Alternative Hypothesis: In major outages caused by severe weather events, there is an association between the specific cause detail and outage duration.

In this hypothesis testing, we repeatedly sampled groups of 181 observations of outages from the overall 669 severe weather related outages without replacement, and compute their average durations. If the null hypothesis is true, it would be common to see an average as high as 4977 minutes. The statistical significant threshold of this test is 0.05 with a test statistic of average outage duration.

Simulation

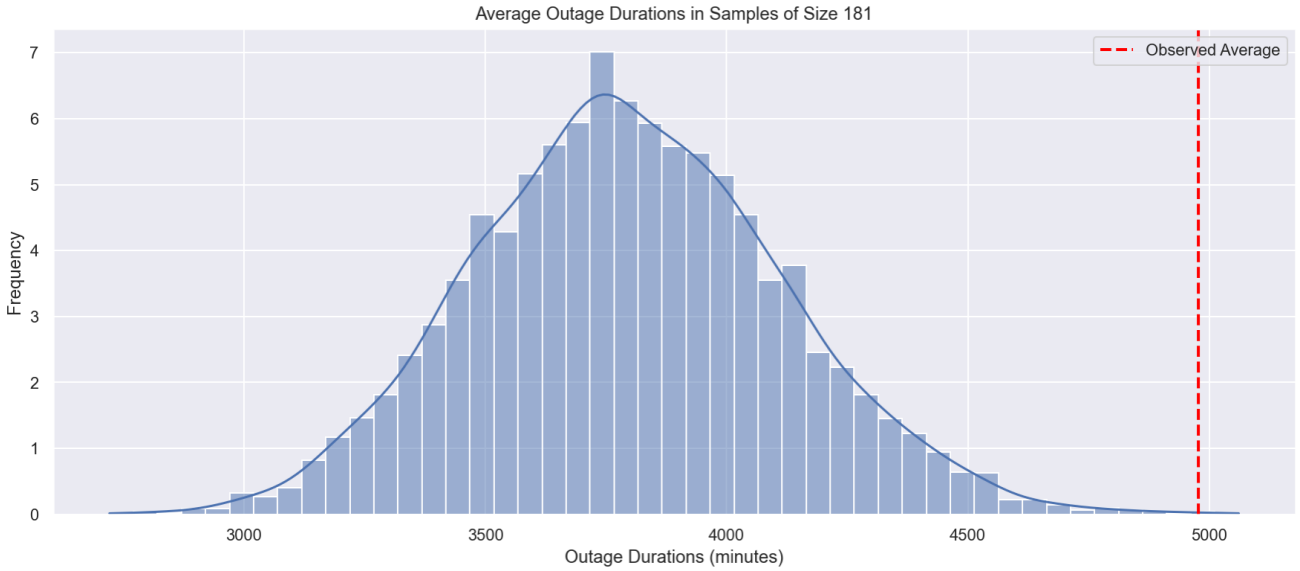
In [21]:
```
num_reps = 5000
num_obs = 181
```

In [22]:
```
avgs = []
avgs = np.nanmean(np.random.choice(weather_df['OUTAGE.DURATION'], size = (num

#visualize simulation result
plt.figure(figsize = (15, 6))
g = sns.histplot(avgs,
                 stat = 'frequency',
                 kde = True)
g.set_xlabel('Outage Durations (minutes)')
g.set_title('Average Outage Durations in Samples of Size 181')

#adds observed test statistic to the graph
observed_average = weather_df.groupby('WIND').mean(numeric_only = True).loc[T
plt.axvline(x = observed_average,
            color='red',
            linewidth=2,
            linestyle = '--',
            label = 'Observed Average');
plt.legend(loc = 'upper right')

plt.show()
```



As shown in by the visualization, the mean durations of outage generated in our simulation has a normal distribution centered around 3800 minutes. This indicates that on avaerga, outage cuased by sereve weather last 3800 minutes. This is significantly less then our observed outage duration which is around 5000 minutes.

In [23]:
```
p_val = (np.array(avgs) >= observed_average).mean() #calculates the p-value o
p_val
```

Out[23]: 0.0006

With 5000 simulations and a p-value < statistical significant threshold (p = 0.05), we reject null hypothesis and conclude that in outage related to severe weather, wind induced power outage has a prolonged duration compared to outage caused by other categories.

Conclusion

In this study, we are trying to find what characteristics might contribute to a major power outage. First, we found that amoung all the major power outages in the past few years, 83.8% of the major outages were induced by severe weather events. We thus narrowed our question down to what are some characteristics of severe weather condition that would lead to a lasting power outage. Based on our knowledge and research on related topic (Hurricane-induced power outage risk under climate change is ...), we hypothesize that major outages caused by heavy wind would result in more prolonged outage comparing to outage caused by other severe weather events.

Before verifying our hypothesis, we first analyzed the missingness of columns that we will be focusing on, `CAUSE.CATEGORY.DETAIL` and conclude that the missingness in this column is MAR. We then decide to drop the rows with null values in the `CAUSE.CATEGORY.DETAIL` since the data contained in this column is categorical and MAR.

Then, with the data cleaned and missingness handled, we performed a hypothesis test to find out whether severe weather with heavy wind would lead to a prolonged power outage. We simulated 5000 samples by resampling the original datatset and conclude that heavy wind does lead to longer durations of power outages with a statistically significant p-value of 0.0006.

Improvements

One thing that could improve our analysis is instead of droping the rows with values in the `CAUSE.CATEGORY.DETAIL` being missing, we could fill in the null values in that columns with plausible data. This is becuase since we already found that the missingness in the 'CAUSE.CATEGORY.DETAIL' column is MAR, ignoring the rows with missing values would lead to a distorted result thus make our analysis potentially biased.

Summary of Findings

Introduction

Power outage is a severe yet common condition that we all have experienced in our life. When a power outage occurs, it could disrupt communications, water and transportation; services as grocery stores, gas stations, ATMs, banks, retail businesses would be paralyzed as well. And when a power outage is prolonged, it could even cause food spoilage and water contamination. Meanwhile, with rising global average temperature and increased emission of greenhouse gas, extreme weather events become more frequent and intense. Would more frequent extreme weather events result in more occurences of power outages? Would specific severe weathers, such as storms, leads to longer duration of outages?

The data used in this project can be found at Major Power Outage Risks. This dataset contains major outages, information on geographical location of the outages, regional climatic information, land-use characteristics, electricity consumption patterns and economic characteristics of the states affected by the outages. It includes major power outages as observations and contains variables such as cause of outage and duration of outage which are helpful to our analysis.

Cleaning and EDA

After importing the raw data, we first found out that the very first several rows and columns are useless. So we removed those rows and columns. Next, We converted the column `OUTAGE.DURATION` to float type for furture usage. After that, we combined the `OUTAGE.START.DATE` and `OUTAGE.START.TIME` columns as well as the `OUTAGE.RESTORATION.DATE` and `OUTAGE.RESTORATION.TIME` columns. We stored the these new columns to columns `OUTAGE.START` and `OUTAGE.RESTORATION` and droped their precursors. Finally, we filtered out power outage that affected at least 50,000 customers or caused an unplanned firm load loss of atleast 300MW according to our definition of major power outages.

Assessment of Missingness

In order to assess the missingness of our data, we first found out that the columns with non-trival missinngness are `HURRICANE.NAMES`, `DEMAND.LOSS.MW`, and `CAUSE.CATEGORY.DETAIL`. Then we decided to analyze the missingness of column `CAUSE.CATEGORY.DETAIL` as it relates to the interest of our study.
When assessing the missingness of this column, we first performed permutation tests between the `CAUSE.CATEGORY` column and our target column. The p-value of this permutation test is 0.0, indicating that the missiningness of data of cause category detail depends on the `CAUSE.CATEGORY` column.
Then, we performed more permutation tests between the target column and columns of `ANOMALY.LEVEL`, `TOTAL.PRICE`, and `TOTAL.SALES` respectively. The p-values for these tests are 0.745, 0.0046, and 0.0008, indicating that the missingness of column `CAUSE.CATEGORY.DETAIL` also depends on the columns of `TOTAL.PRICE` and `TOTAL.SALES` but not column `ANOMALY.LEVEL`.

Thus, we conclude that the missingness in the `CAUSE.CATEGORY.DETAIL` column is MAR (missing at random). This means that ignoring the missingness or imputing the missing with a designated value would very likely lead to bias.

Hypothesis Test

In this hypothesis test of the effect outage cause details on outage duration, we want to verify whether there is an association between the specific cause detail and outage duration. In order to do so, we performed a hypothesis test by repeatedly sampled groups of 181 observations of outages from the overall 669 severe weather related outages without replacement, and compute their average durations. If there is no association, it would be common to see an average as high as 4977 minutes. We set the statistical significant threshold of this test as 0.05 with a test statistic of average outage duration. Our result of p = 0.0006 suggests that there is indeed an association between the specific cause detail and outage duration, specifically, outage induced by heavy wind events have a prolonged duration compared to outages caused by other severe weather events.

In [ ]: