# Show delays for all reasons in the airports identified with at least 1 month >50 weather delays in 2003 to 2022.

# Data are from 2003 to 2022

# Data are sorted in descending order showing airport with most delays in 20 years first

group\_AllAirports\_stat = group\_AllAirports\_stat.sort\_values(by = ['Delayed'], ascending = False)

group\_AllAirports\_stat

# Tried to find the maximum from each column but did not work

# Can someone please help?

col\_max = group\_AllAirports\_stat.max(axis=0)

col\_max

# Calculate summary statistics for each airport for the 5 year period 1028 to 2022 period.

# Can someone help here? It will not limit to the last 5 years.

#Refine Data for years b/w 2018 and 2022

#group\_AllAirports\_18to22 = group\_AllAirports\_stat[(group\_AllAirports\_stat["year"]== 2022)]

#(group\_AllAirports\_stat["Year"] <= 2022)

#]

group\_AllAirports\_stat.columns

# group\_AllAirports\_18to22\_stat = group\_AllAirports\_18to22.groupby("airport")

# summary\_Airports\_stat = group\_AllAirports\_18to22\_stat['arr\_del15'].agg([np.mean, np.median, np.var, np.std, pd.Series.sem])

# summary\_Airports\_stat = summary\_Airports\_stat.sort\_values(by = ['mean'], ascending = False)

# #summary\_AllAirlines\_allyears\_stat.rename(columns={'mean': 'Mean',

# # 'median': 'Median',

# # 'var': 'Variance',

# # 'std': 'Standard Deviation',

# # 'sem': 'Standard Error of the Mean'}, inplace=True)

# print(summary\_Airports\_allyears\_stat)

# This shows the average number of dealys for any reason at all airports per year in US

# \*\*\* FOR PRESENTATION \*\*\*

# What is the average number of delayed flights each year at each airport? Plot the year, mean and SD.

# Insight: Over the past 20 years, highest average delays are in Atlanta, Chicago, Dallas, Los Angeles and Denver.

# DO NOT USE

# What is the average number of delayed flights per month at each airport? Plot the year, mean and SD.

# Calculate summary statistics for each airport for the whole 20-year period.

group\_AllAirports\_stat = df5\_AirportDelay\_2003to2022.groupby(["Airport"]).sum()

summary\_Airports\_allyears\_stat = group\_AllAirports\_stat['arr\_del15'].agg([np.mean, np.median, np.var, np.std, pd.Series.sem])

summary\_Airports\_allyears\_stat = summary\_Airports\_allyears\_stat.sort\_values(by = ['mean'], ascending = False)

# #summary\_AllAirlines\_allyears\_stat.rename(columns={'mean': 'Mean',

# # 'median': 'Median',

# # 'var': 'Variance',

# # 'std': 'Standard Deviation',

# # 'sem': 'Standard Error of the Mean'}, inplace=True)

print(summary\_Airports\_allyears\_stat)

# This shows the average number of dealys for any reason at all airports per year in US

# \*\*\* FOR PRESENTATION \*\*\*

# Insight: Over the past 20 years, highest average delays are in Atlanta, Chicago, Dallas, Los Angeles and Denver.

### Calculate summary statistics for each airline for 2022.

df\_AirlineDelayCleanDS18to22 = df\_AirlineDelayCleanDS[(df\_AirlineDelayCleanDS["Year"]>= 2018) &

(df\_AirlineDelayCleanDS["Year"] <= 2022)

]

group\_AllAirlines\_stat = df\_AirlineDelayCleanDS18to22.groupby("carrier\_name")

summary\_AllAirlines\_5years\_stat = group\_AllAirlines\_stat['arr\_del15'].agg([np.mean, np.median, np.var, np.std, pd.Series.sem])

summary\_AllAirlines\_5years\_stat = summary\_AllAirlines\_5years\_stat.sort\_values(by = ['mean'], ascending = False)

#summary\_AllAirlines\_allyears\_stat.rename(columns={'mean': 'Mean',

# 'median': 'Median',

# 'var': 'Variance',

# 'std': 'Standard Deviation',

# 'sem': 'Standard Error of the Mean'}, inplace=True)

print(summary\_AllAirlines\_allyears\_stat)###

# Use this to show how we identified the airlines with the most delays over the 20-year period.

# What is the average number of delayed flights for each airline over all 20 years

# Not required, plotted in excel from Table 7

# Create dataframe for percent of delay from each reason for 2022 only

df2\_DelayReasons22 = df2\_AirlineDelay\_AllDelays\_Rollup[(df2\_AirlineDelay\_AllDelays\_Rollup["Year"]== 2022)]

df2\_DelayReasons22

#df2\_DelayReasons22\_refined = df2\_DelayReasons22[["Year","airport","arr\_flights","arr\_del15","% Delayed d/t Airline","% Delayed d/t Weather","% Delayed d/t NAS","% Delayed d/t Security","% Delayed d/t Late Aircraft"]]

#df3\_DelayReasons22graph = pd.df2\_DelayReasons22\_refined({"year",index=[''% Delayed d/t Airline','% Delayed d/t Weather',"% Delayed d/t NAS","% Delayed d/t Security","% Delayed d/t Late Aircraft"])

#plot = df3\_DelayReasons22graph.plot.pie(y='year', figsize=(5, 5))

# CREATE A PIE CHART of the data from 2022

# Include the data from the last 5 columns (from % d/t airine to % due to aircraft)

# Show the plot

plt.show()

# INSIGHTS - discuss how each contributes to the overall amount of delayed flights.

# Say that 333,916 (19.01%) flights were delayed of all 1,756,855 arriving flights in 2022.

# Delays due to weather were a small contributor to the overall delays (3.45%). THe most delays were due to the airline (42.24), eg, availablity of crew).

# Used Excel, delete this block

# CREATE A bar chart of the above data for the 5 years and 5 airlines

df\_AirlineCarrierDelay\_PerYear

Airline\_Year\_string = ["2018","2019","2020","2021","2022"]

AirlineYears = [int(i) for i in Airline\_Year\_string]

hd = df2\_AirlineDelayCleanDS\_Status\_Rollup["arr\_del15"]

y\_axis = np.arange(len(hd))

plt.bar(AirlineYear,hd,width=0.8, bottom=None, color='y', align='center')

# Include plot title and axis labels

plt.title("Number of Delayed Arriving Flights in US (2003 to 2022)")

plt.xlabel("Year")

plt.ylabel("Percent of Flights Delayed")

plt.xticks(range(2003,2023))

plt.xticks(rotation=45)

plt.show()

#Total minutes of delays in all US for delayed flights for 5 years any reason for 5 airline carriers

df\_AirlineDelayTotalMins5Carriers = df\_AirlineDelayGreater50WeatherCt[["Year","carrier\_name","arr\_delay","carrier\_delay","weather\_delay", "nas\_delay",

"security\_delay","late\_aircraft\_delay"

]]

df\_AirlineDelayTotalMins5Carriers.sort\_values(by = ['arr\_delay'], ascending = False)

#Airline Carriers that had the highest delays for each year

df\_AirlineDelayTotalMins5CarrMaxYr =df\_AirlineDelayTotalMins5Carriers.groupby(df\_AirlineDelayTotalMins5Carriers['year']).max()

df\_AirlineDelayTotalMins5CarrMaxYr

df\_AirlineDelayArrConvertToHrs= df\_AirlineDelayTotalMins5CarrMaxYr['arr\_delay']/60

df\_AirlineDelayCarrConvertToHrs= df\_AirlineDelayTotalMins5CarrMaxYr['carrier\_delay']/60

df\_AirlineDelayWeatherConvertToHrs= df\_AirlineDelayTotalMins5CarrMaxYr['weather\_delay']/60

df\_AirlineDelayNASConvertToHrs= df\_AirlineDelayTotalMins5CarrMaxYr['nas\_delay']/60

df\_AirlineDelayTotalMins5CarrMaxYr["arr\_delay\_hrs"] = df\_AirlineDelayArrConvertToHrs

df\_AirlineDelayTotalMins5CarrMaxYr["carrier\_delay\_hrs"] = df\_AirlineDelayCarrConvertToHrs

df\_AirlineDelayTotalMins5CarrMaxYr["weather\_delay\_hrs"] = df\_AirlineDelayWeatherConvertToHrs

df\_AirlineDelayTotalMins5CarrMaxYr

#

df\_AirlineDelayGreater50WeatherCt.groupby("month").max()["weather\_ct"]

df\_AirlineDelayGreater50WeatherCt.max(axis=0)

#df\_AirlineDelayGreater50WeatherCt.groupby(["carrier\_name","Year","month"]).sum().head(20)

df\_AirlineDelayGreater50WeatherCt.groupby(["carrier\_name","Year","month"]).sum().tail(20)

df\_AirlineDelayGreater50WeatherCt.to\_csv(AirlineDelay\_CleanOutput)