# Flight Delay Claim Analysis

## Insights and recommendations for better pricing decisions.

Santiago Codaro

04/23/2022

# Objective

The objective of this presentation is to shed light on the factors that attribute to a “low-risk” or “high-risk” flight so insurance prices can be set accordingly.

Furthermore, a classification rule is going to be proposed to label the flights by their risk level.

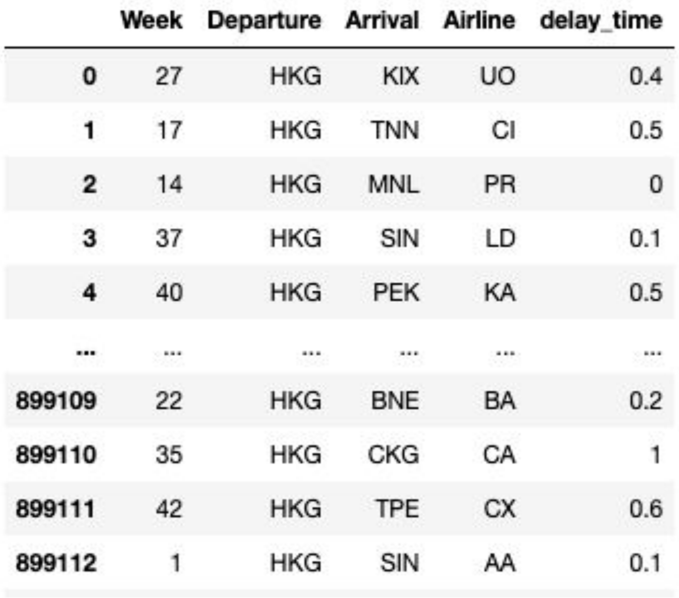
# Agenda

- Logistic Regression Model. go
- Flight Delay Deep Dive. go
- Conclusions. go
- Appendix. go

# Logistic Regression Model

# The Model

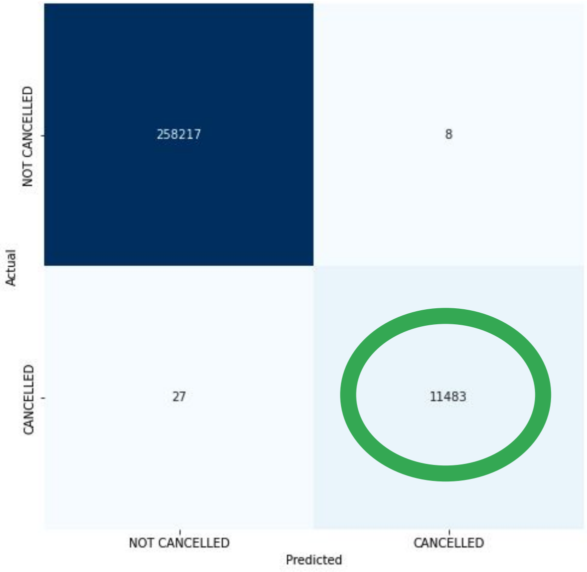
Logistic regression models are used to predict yes/no outcomes based on a given set of variables. In this case, we are predicting if there was a claim or not, based on the following variables:



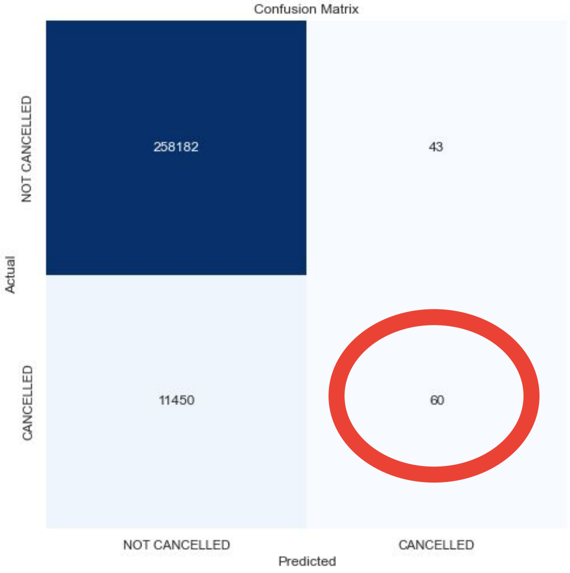
| | Week | Departure | Arrival | Airline | delay_time |
|---|---|---|---|---|---|
| 0 | 27 | HKG | KIX | UO | 0.4 |
| 1 | 17 | HKG | TNN | CI | 0.5 |
| 2 | 14 | HKG | MNL | PR | 0 |
| 3 | 37 | HKG | SIN | LD | 0.1 |
| 4 | 40 | HKG | PEK | KA | 0.5 |
| ... | ... | ... | ... | ... | ... |
| 899109 | 22 | HKG | BNE | BA | 0.2 |
| 899110 | 35 | HKG | CKG | CA | 1 |
| 899111 | 42 | HKG | TPE | CX | 0.6 |
| 899112 | 1 | HKG | SIN | AA | 0.1 |

# Model Results

Including delay column



Excluding delay column



When the ‘delay_time’ column in **included**, the model can predict well up to 99% of the cancelled flights. However, if we **exclude** it, the model can’t predict almost none of the actually cancelled flights. This means that…

> “The only attribute that can predict flight claims is its historical flight delay.”

Now that we know this, let’s deep dive into the flight delay data:

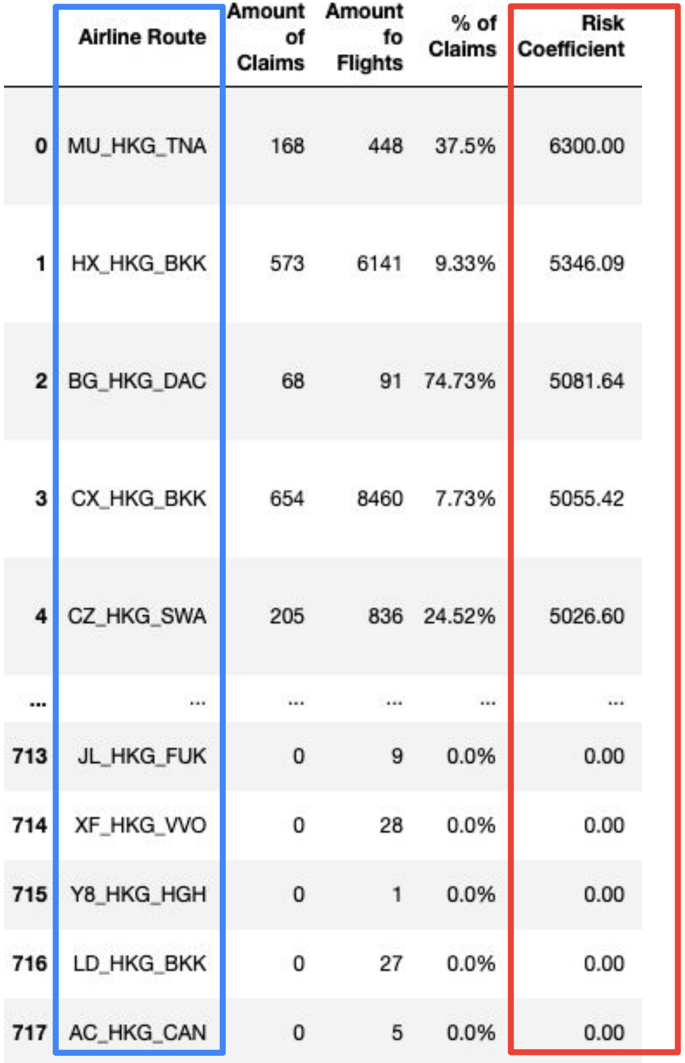
# Flight Delay Deep Dive

# Some Info First

Data was grouped by the combination of 3 dimensions of each flight: The Airline, the Departure Airport and its Arrival Airport (a.k.a '**Airline Route**')

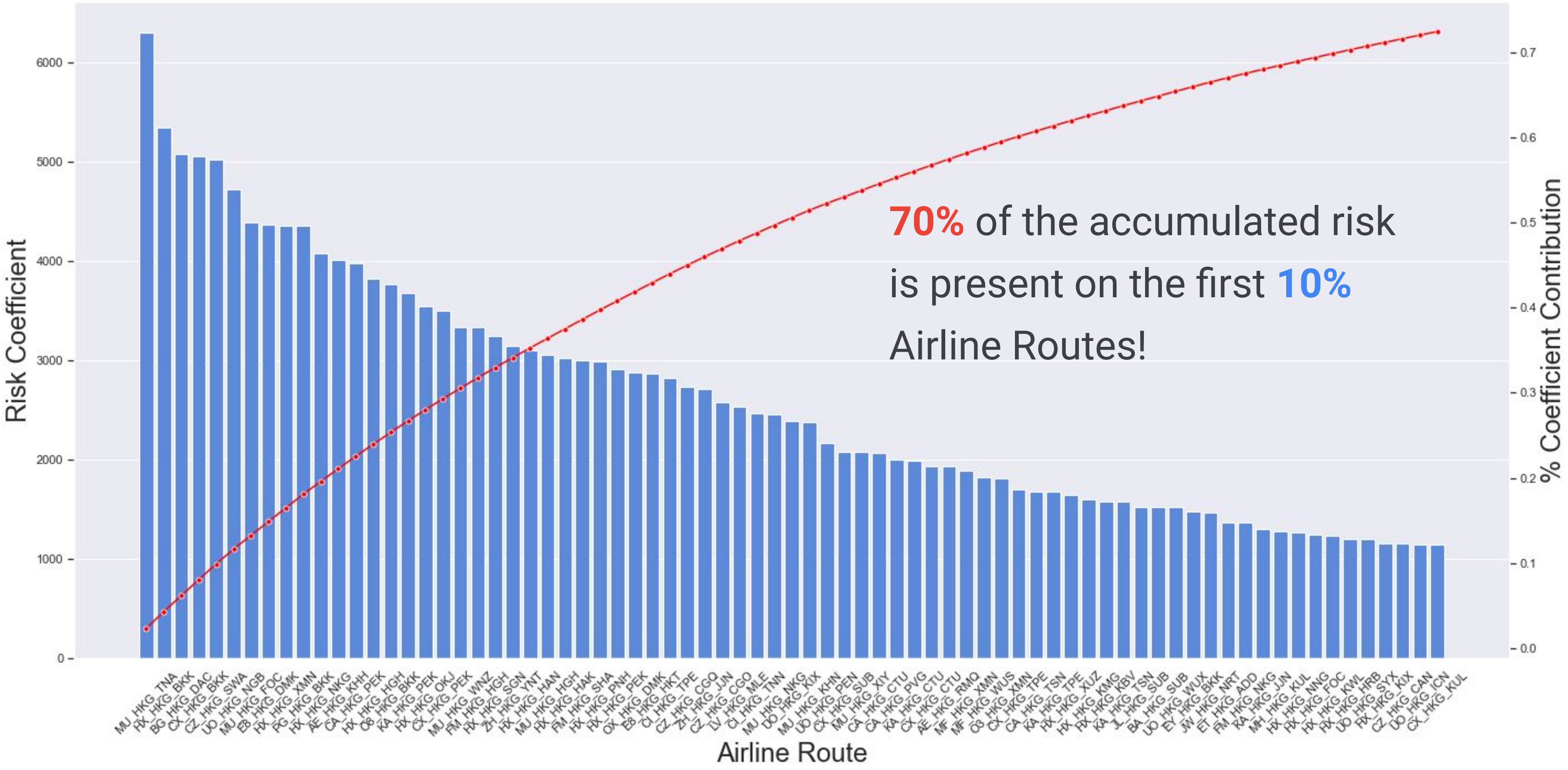
The column '**Risk Coefficient**'* weights both the % of claims and the amount of claims. Highest values mean riskier Airline Routes.

* Risk Coefficient = (% of claims * 100) * Amount of Claims



| | Airline Route | Amount of Claims | Amount fo Flights | % of Claims | Risk Coefficient |
|---|---|---|---|---|---|
| 0 | MU_HKG_TNA | 168 | 448 | 37.5% | 6300.00 |
| 1 | HX_HKG_BKK | 573 | 6141 | 9.33% | 5346.09 |
| 2 | BG_HKG_DAC | 68 | 91 | 74.73% | 5081.64 |
| 3 | CX_HKG_BKK | 654 | 8460 | 7.73% | 5055.42 |
| 4 | CZ_HKG_SWA | 205 | 836 | 24.52% | 5026.60 |
| ... | ... | ... | ... | ... | ... |
| 713 | JL_HKG_FUK | 0 | 9 | 0.0% | 0.00 |
| 714 | XF_HKG_VVO | 0 | 28 | 0.0% | 0.00 |
| 715 | Y8_HKG_HGH | 0 | 1 | 0.0% | 0.00 |
| 716 | LD_HKG_BKK | 0 | 27 | 0.0% | 0.00 |
| 717 | AC_HKG_CAN | 0 | 5 | 0.0% | 0.00 |

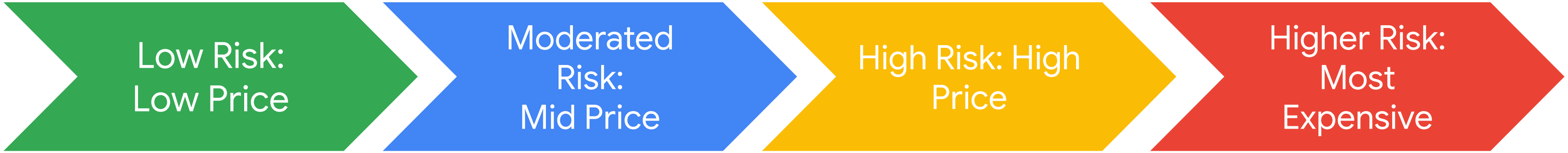
# Pareto's Rule



**70%** of the accumulated risk is present on the first **10%** Airline Routes!

# Categories

To identify the risk of each flight, each Airline-Route was labeled accordingly to its accumulated Risk Coefficient:



Low Risk: Low Price

Moderated Risk: Mid Price

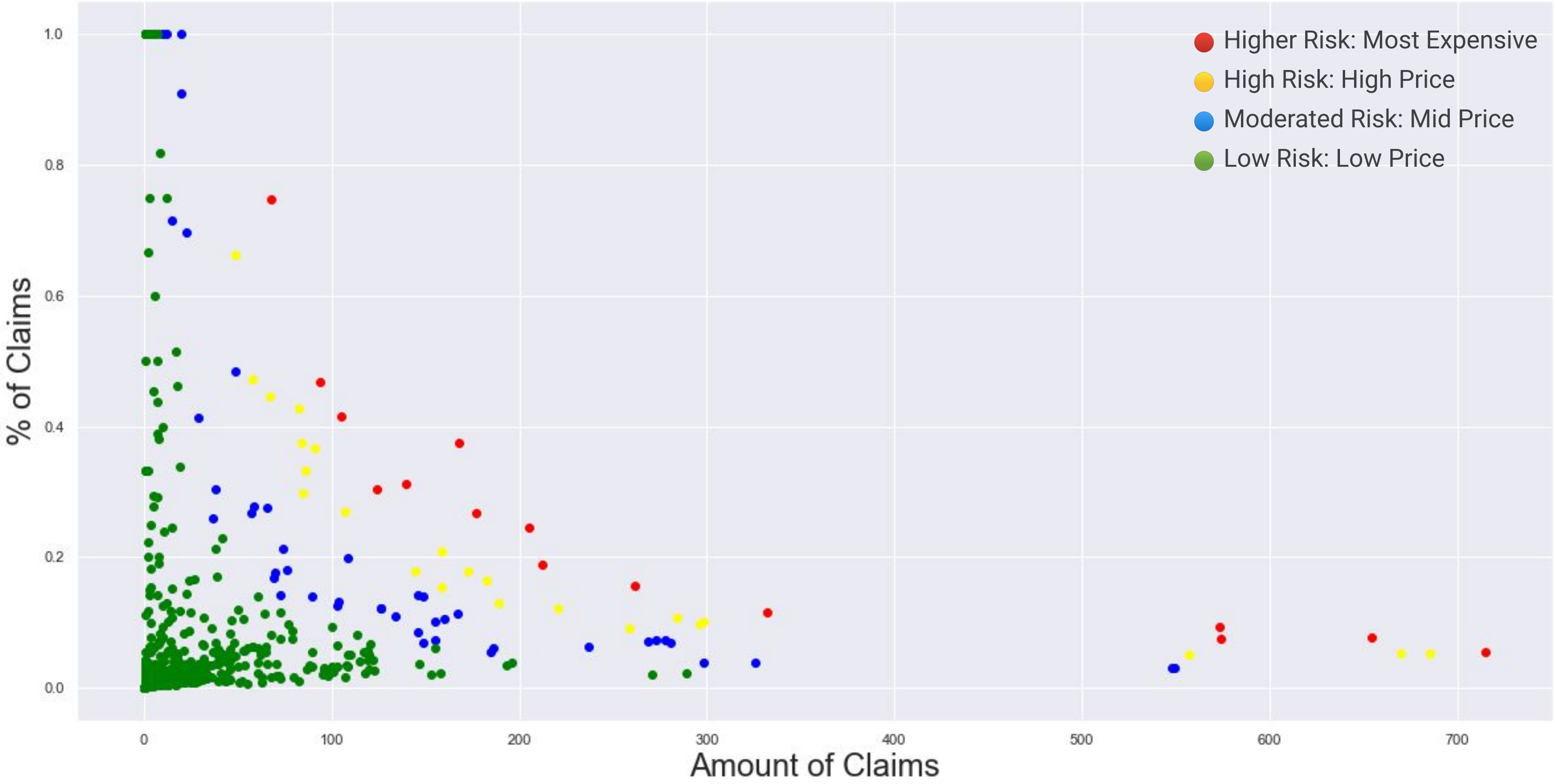
High Risk: High Price

Higher Risk: Most Expensive

| | | | |
|---|---|---|---|
| 25% | 50% | 75% | 100% |
| 958 rc | 2376 rc | 3678 rc | 6300 rc |

Accumulated Risk Coefficient(rc) →

# Risk Categories Distribution



# Conclusions

# Conclusions and Recommendations

📈 During the analysis, a very strong correlation between flight delay and claim requests was found while the correlation among other variables was not significative.

🛩️ Based on this finding, a risk coefficient was created. Then, each Airline-Route was labeled according to its accumulated % contribution to the mentioned coefficient.

💵 It is recommended to set prices according to the categories. For future unlabeled Airline-Routes, the claim risk coefficient moving averages of the last 8 weeks could be used to determine the risk category.

# Thanks!

You can find the full technical analysis here.