

#TidyTuesday

Week 39 - Himalayan Climbing Expeditions

TheDataAreClean

27th September, 2020

Contents

Setup	1
R-Markdown	2
Analysis	2
Data	2
Read	2
Explore	3
Distribution	4
Over Time	7
One Last Thing	12
Questions	14
Next Steps	14

The data this week comes from The Himalayan Database.

The Himalayan Database is a compilation of records for all expeditions that have climbed in the Nepal Himalaya. The database is based on the expedition archives of Elizabeth Hawley, a longtime journalist based in Kathmandu, and it is supplemented by information gathered from books, alpine journals and correspondence with Himalayan climbers.

The data cover all expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. Also included are expeditions to both sides of border peaks such as Everest, Cho Oyu, Makalu and Kangchenjunga as well as to some smaller border peaks. Data on expeditions to trekking peaks are included for early attempts, first ascents and major accidents.

Setup

To analyse the data, first step is to setup the working environment.

R-Markdown

Setting local settings for R-Markdown output characteristics.

Analysis

Loading all the required libraries to complete the analysis.

```
# Load required packages
library(tidyuesdayR)
library(tidyverse)
```

```
## -- Attaching packages ----- tidy

## v ggplot2 3.3.2      v purrr  0.3.4
## v tibble  3.0.3      v dplyr  1.0.0
## v tidyr   1.1.0      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0

## -- Conflicts ----- tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data

Read

Import the data from the tidyuesdayR package and store separate files.

```
# Load the #TidyTuesday dataset
## Read Week 39 Data
tuesdata <- tidyuesdayR::tt_load(2020, week = 39)
```

```
## --- Compiling #TidyTuesday Information for 2020-09-22 ----

## --- There are 3 files available ---

## --- Starting Download ---

##
## Downloading file 1 of 3: 'peaks.csv'
## Downloading file 2 of 3: 'members.csv'
## Downloading file 3 of 3: 'expeditions.csv'

## --- Download complete ---
```

```
## Store individual datasets
## Peaks
peaks <- tuesdata$peaks

## Members
members <- tuesdata$members

## Expeditions
expeditions <- tuesdata$expeditions
```

Explore

Preliminary analysis of the datasets.

Looking at the peaks data.

```
# Review Peaks Dataset
## Structure
glimpse(peaks)

## Rows: 468
## Columns: 8
## $ peak_id          <chr> "AMAD", "AMPG", "ANN1", "ANN2", "ANN3", ...
## $ peak_name        <chr> "Ama Dablam", "Amphu Gyabjen", "Annapurn...
## $ peak_alternative_name <chr> "Amai Dablang", NA, NA, NA, NA, NA, ...
## $ height_metres    <dbl> 6814, 5630, 8091, 7937, 7555, 7525, 8026...
## $ climbing_status  <chr> "Climbed", "Climbed", "Climbed", "Climbe...
## $ first_ascent_year <dbl> 1961, 1953, 1950, 1960, 1961, 1955, 1974...
## $ first_ascent_country <chr> "New Zealand, USA, UK", "UK", "France", ...
## $ first_ascent_expedition_id <chr> "AMAD61101", "AMPG53101", "ANN150101", "...
```

```
## Peaks Climbed
peaks %>%
  group_by(climbing_status) %>%
  summarise(count = n())
```

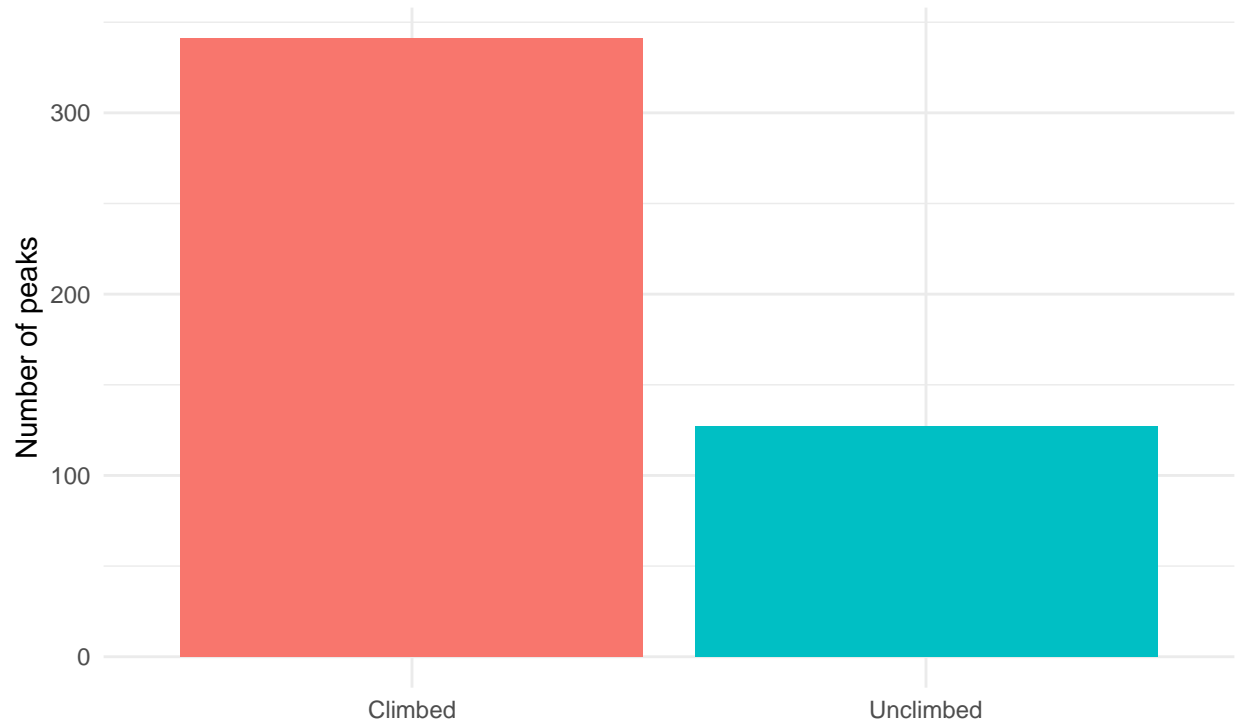
```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
## # A tibble: 2 x 2
##   climbing_status count
##   <chr>           <int>
## 1 Climbed         341
## 2 Unclimbed       127
```

```
## Peaks Climbed Bar Chart
peaks %>%
  ggplot(aes(climbing_status, fill = climbing_status)) +
  geom_bar() + theme_minimal() +
  labs(
    title = "More than a quarter of Himalayan peaks remain unclimbed",
    caption = "Source: The Himalayan Database",
```

```
x = "",
y = "Number of peaks"
) +
theme(legend.position = "none")
```

More than a quarter of Himalayan peaks remain unclimbed



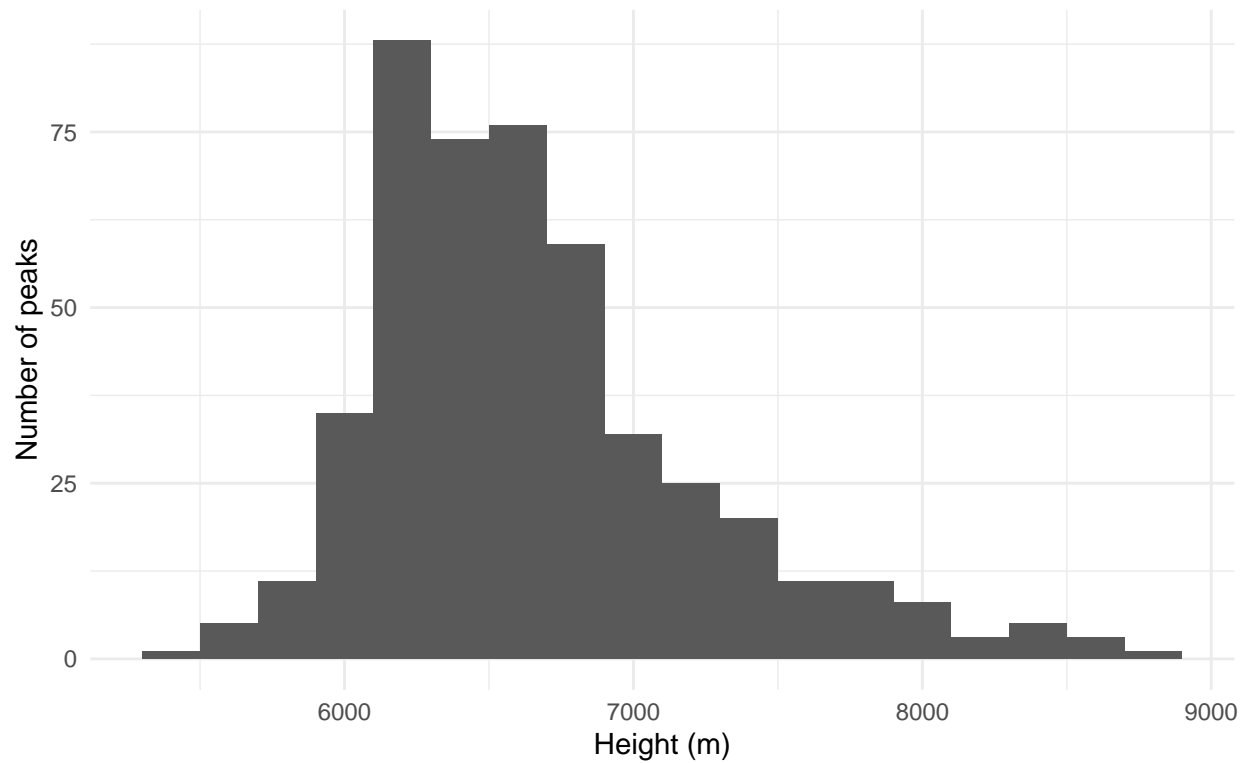
Source: The Himalayan Database

Distribution

Next, I wanted to look at distribution of the peak heights.

```
## Peaks and Heights
peaks %>%
  ggplot(aes(x = height_metres)) +
  geom_histogram(binwidth = 200) + theme_minimal() +
  labs(
    title = "How tall are Himalayan peaks?",
    caption = "Source: The Himalayan Database",
    x = "Height (m)",
    y = "Number of peaks"
  )
```

How tall are Himalayan peaks?

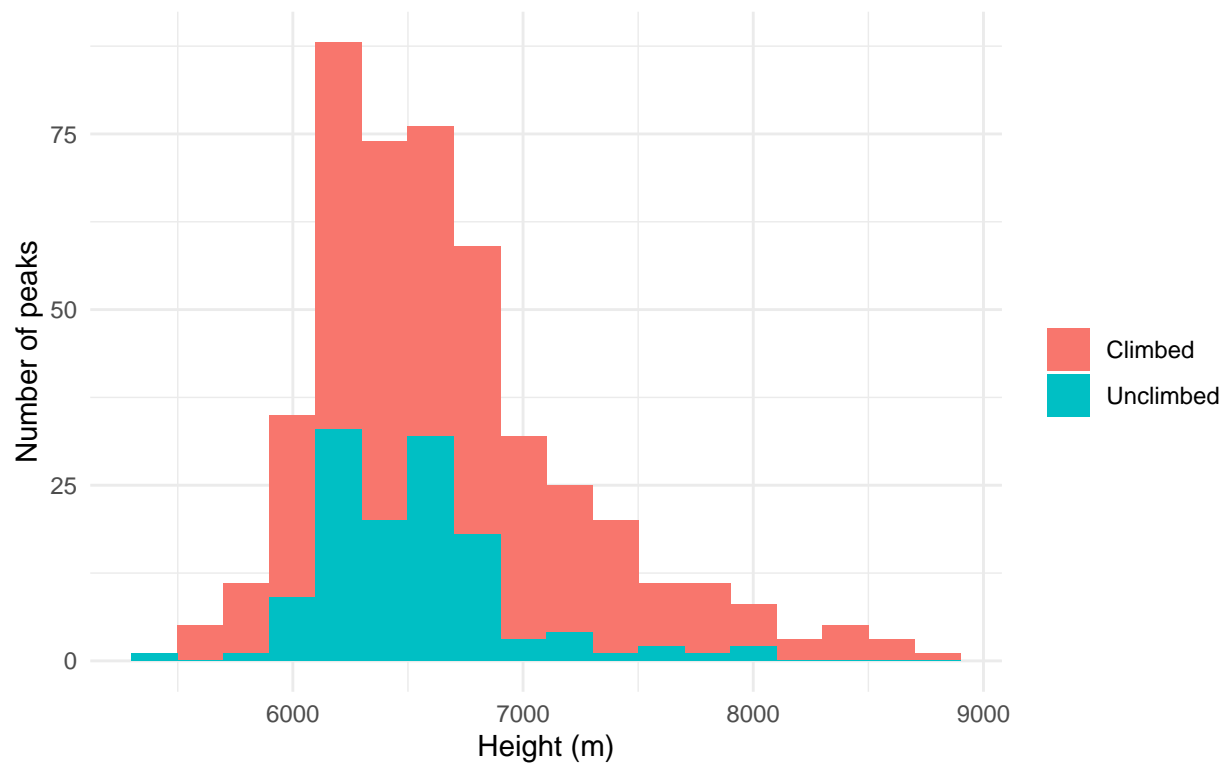


Source: The Himalayan Database

And how many of those peaks are “Unclimbed”.

```
## Peaks, Heights & Status
peaks %>%
  ggplot(aes(x = height_metres, fill = climbing_status)) +
  geom_histogram(binwidth = 200) + theme_minimal() +
  labs(
    title = "How many of the tall peaks are climbed?",
    caption = "Source: The Himalayan Database",
    x = "Height (m)",
    y = "Number of peaks",
    fill = ""
  )
```

How many of the tall peaks are climbed?

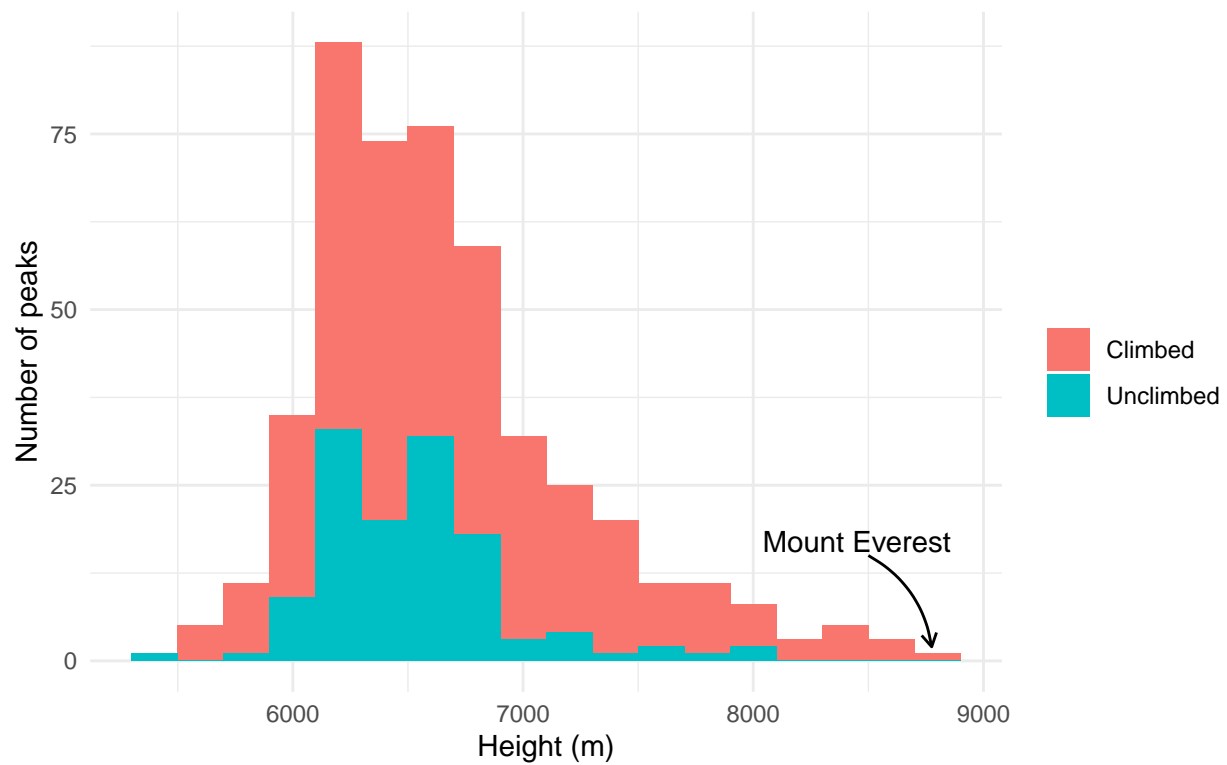


Source: The Himalayan Database

And identifying Mount Everest !

```
## Peaks, Heights, Status and Everest
peaks %>%
  ggplot(aes(x = height_metres, fill = climbing_status)) +
  geom_histogram(binwidth = 200) + theme_minimal() +
  labs(
    title = "What about Mount Everest !",
    caption = "Source: The Himalayan Database",
    x = "Height (m)",
    y = "Number of peaks",
    fill = ""
  ) +
  annotate("text", 8450, 17, label = "Mount Everest") +
  annotate(
    "curve",
    x = 8500,
    y = 15,
    xend = 8775,
    yend = 2,
    curvature = -0.25,
    arrow = arrow(length = unit(2, "mm"))
  )
```

What about Mount Everest !



Source: The Himalayan Database

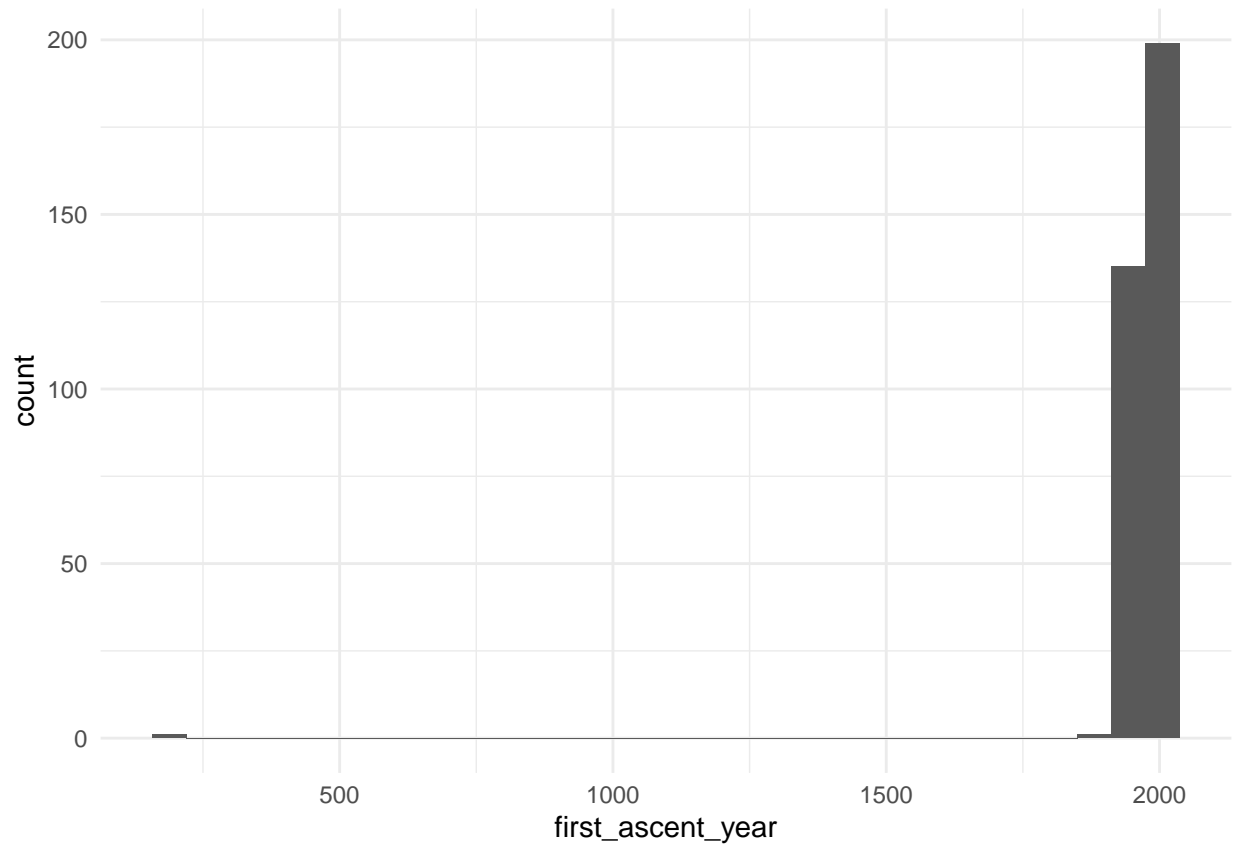
Over Time

When were Himalayan peaks first climbed?

```
## Peaks climbed over years Histogram
peaks %>%
  ggplot(aes(first_ascent_year)) +
  geom_histogram() + theme_minimal()
```

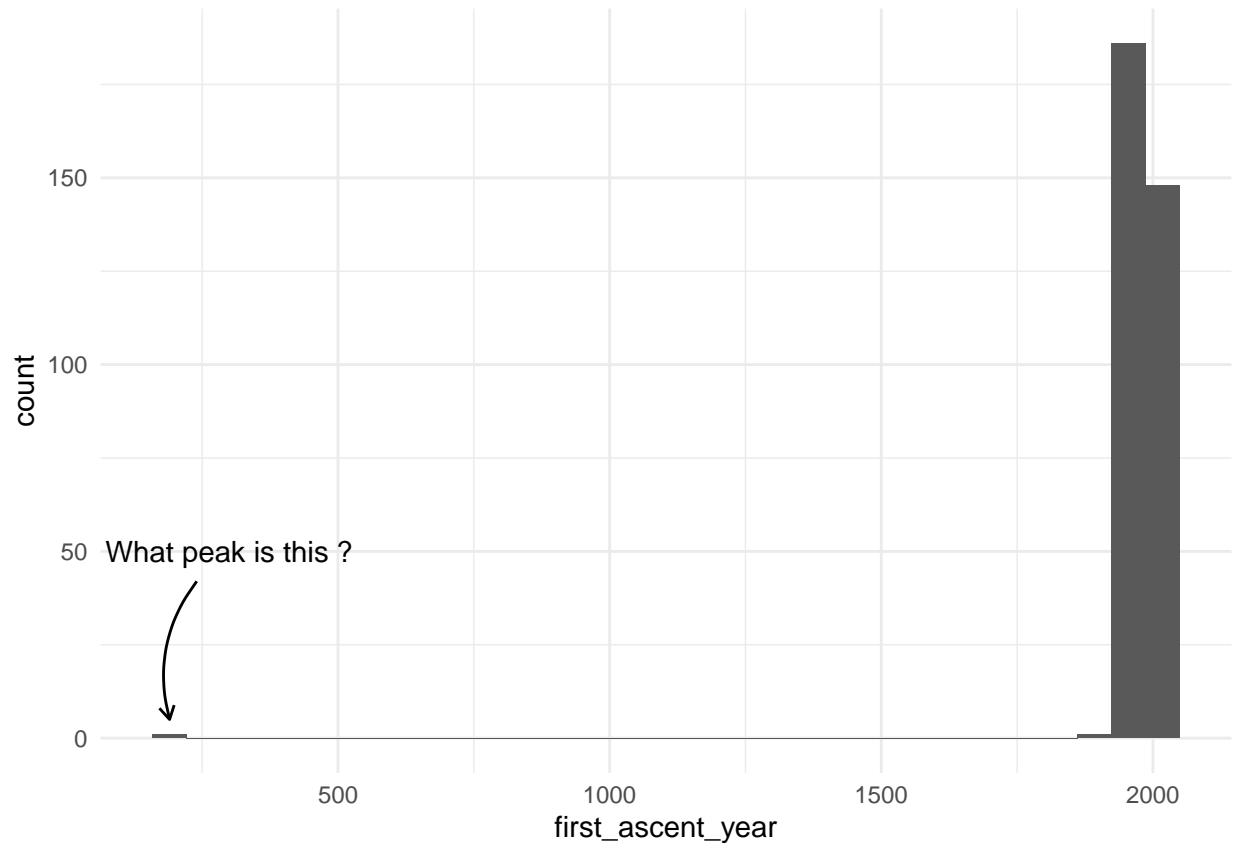
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## Warning: Removed 132 rows containing non-finite values (stat_bin).
```



Identifying the error year and re-visualizing the data.

```
## Identify Anomaly Peak
## Look at the chart
peaks %>%
  ggplot(aes(first_ascent_year)) +
  geom_histogram() + theme_minimal() +
  annotate("text", 300, 50, label = "What peak is this ?") +
  annotate(
    "curve",
    x = 240,
    y = 42,
    xend = 190,
    yend = 5,
    curvature = 0.25,
    arrow = arrow(length = unit(2, "mm"))
  )
```

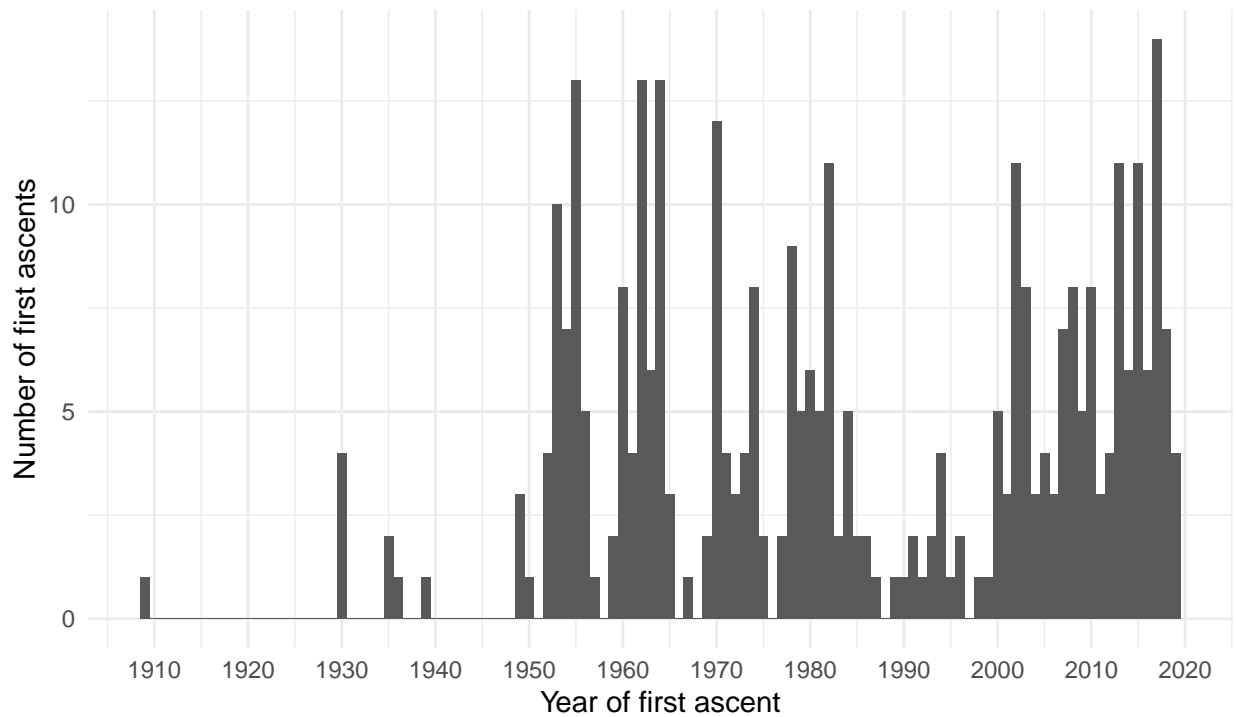
```
## Search Data
## Min Climb Year
min(peaks$first_ascent_year, na.rm = TRUE)
```

```
## [1] 201
```

```
## Filter Out and Visualize
## One Year Buckets
peaks %>%
  filter(first_ascent_year != 201) %>%
  ggplot(aes(first_ascent_year)) +
  geom_histogram(binwidth = 1) + theme_minimal() +
  scale_x_continuous(breaks = seq(1910, 2020, 10)) +
  labs(title = "Climbers are still sumitting peaks for the first time",
       subtitle = "Year of first ascent for Himalayan peaks",
       caption = "Source: The Himalayan Database",
       x = "Year of first ascent",
       y = "Number of first ascents")
```

Climbers are still summitting peaks for the first time

Year of first ascent for Himalayan peaks

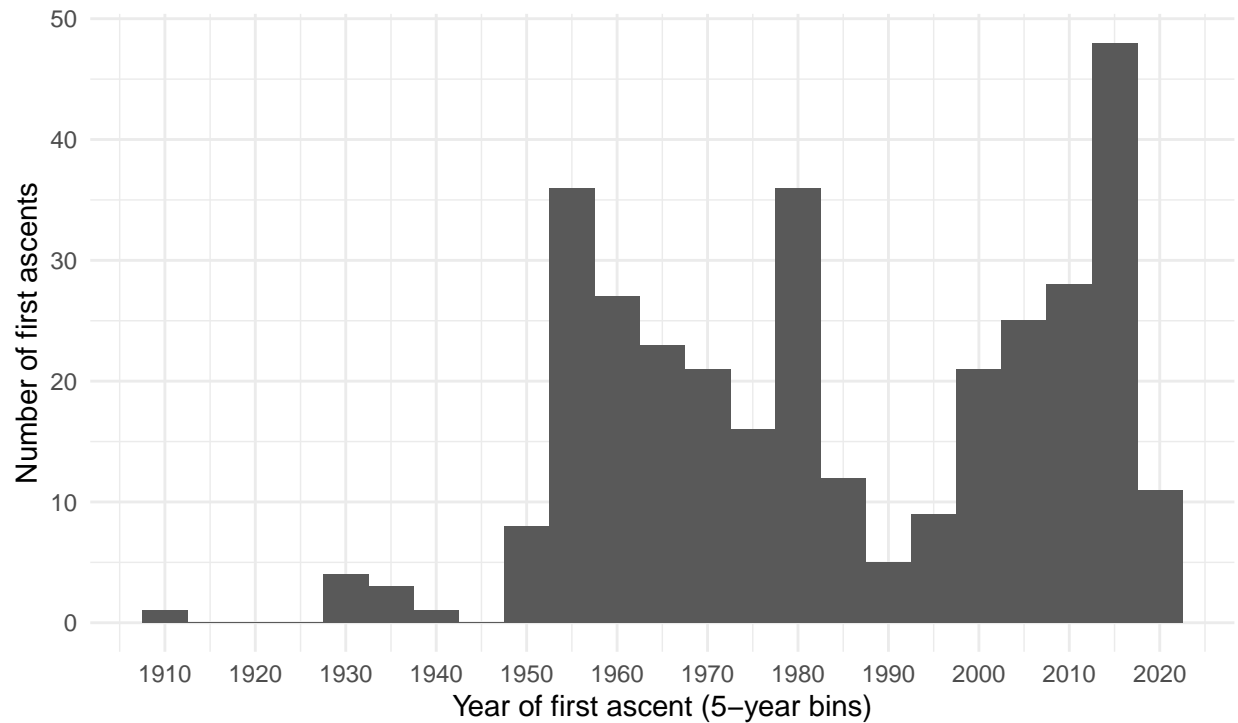


Source: The Himalayan Database

```
## Filter Out and Visualize
## Five Year Buckets
peaks %>%
  filter(first_ascent_year != 201) %>%
  ggplot(aes(first_ascent_year)) +
  geom_histogram(binwidth = 5) + theme_minimal() +
  scale_x_continuous(breaks = seq(1910, 2020, 10)) +
  labs(title = "Climbers are still summitting peaks for the first time",
       subtitle = "Year of first ascent for Himalayan peaks",
       caption = "Source: The Himalayan Database",
       x = "Year of first ascent (5-year bins)",
       y = "Number of first ascents")
```

Climbers are still summitting peaks for the first time

Year of first ascent for Himalayan peaks

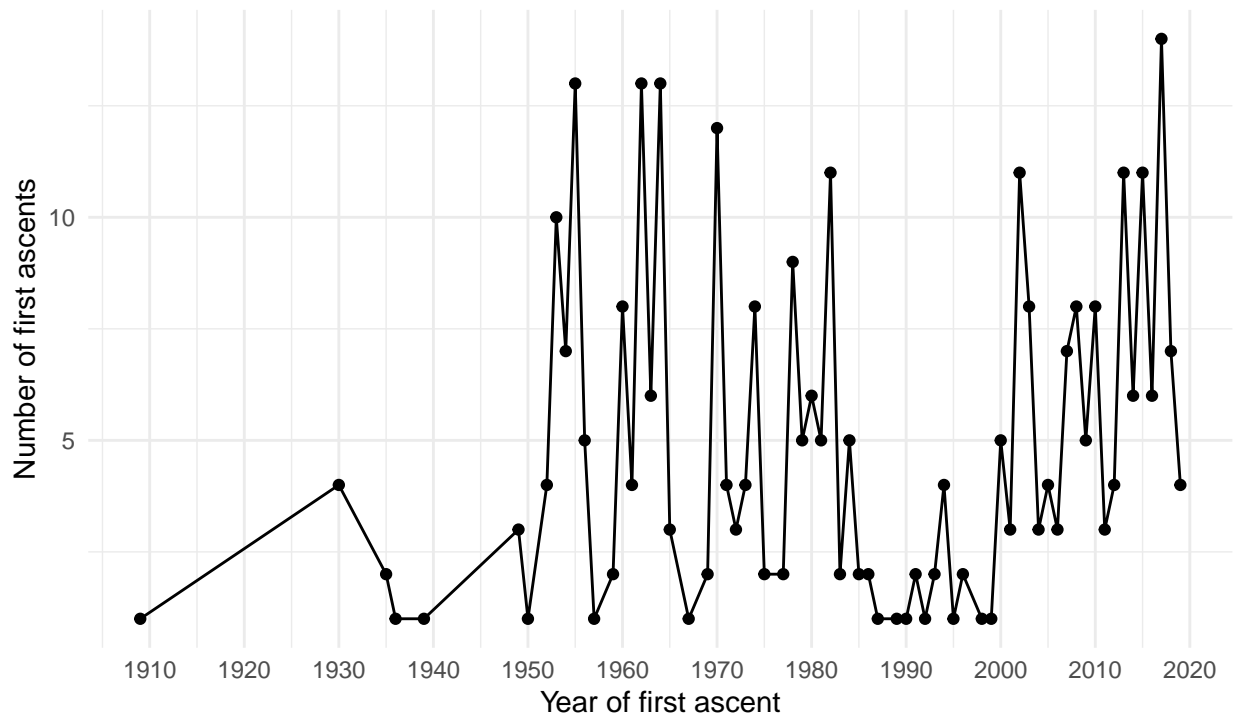


Source: The Himalayan Database

```
## Filter Out and Visualize
## Line Chart + Extra Effort
peaks %>%
  filter(first_ascent_year != 201 & !is.na(first_ascent_year)) %>%
  group_by(first_ascent_year) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = first_ascent_year, y = count)) +
  geom_line() + geom_point() + theme_minimal() +
  scale_x_continuous(breaks = seq(1910, 2020, 10)) +
  labs(title = "Climbers are still summitting peaks for the first time",
       subtitle = "Year of first ascent for Himalayan peaks",
       caption = "Source: The Himalayan Database",
       x = "Year of first ascent",
       y = "Number of first ascents")
```

Climbers are still summitting peaks for the first time

Year of first ascent for Himalayan peaks



Source: The Himalayan Database

The line chart option is considered to show trends over time, doesn't seem to work well in this case.

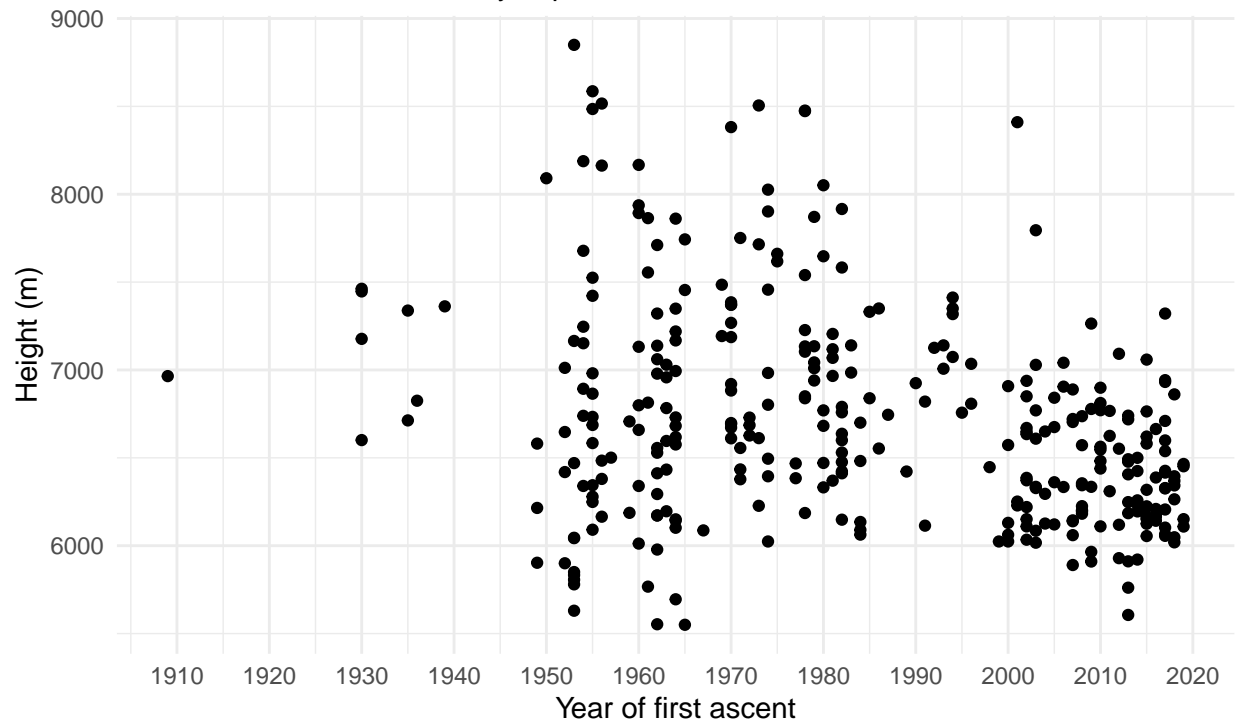
One Last Thing

When did we climb Everest during all of this?

```
# Every over time
peaks %>%
  filter(first_ascent_year != 201 & !is.na(first_ascent_year)) %>%
  ggplot(aes(x = first_ascent_year, y = height_metres)) +
  geom_point() + theme_minimal() +
  scale_x_continuous(breaks = seq(1910, 2020, 10)) +
  labs(title = "What is the human mind thinking ?",
        subtitle = "Year of first ascent for Himalayan peaks",
        caption = "Source: The Himalayan Database",
        x = "Year of first ascent",
        y = "Height (m)")
```

What is the human mind thinking ?

Year of first ascent for Himalayan peaks

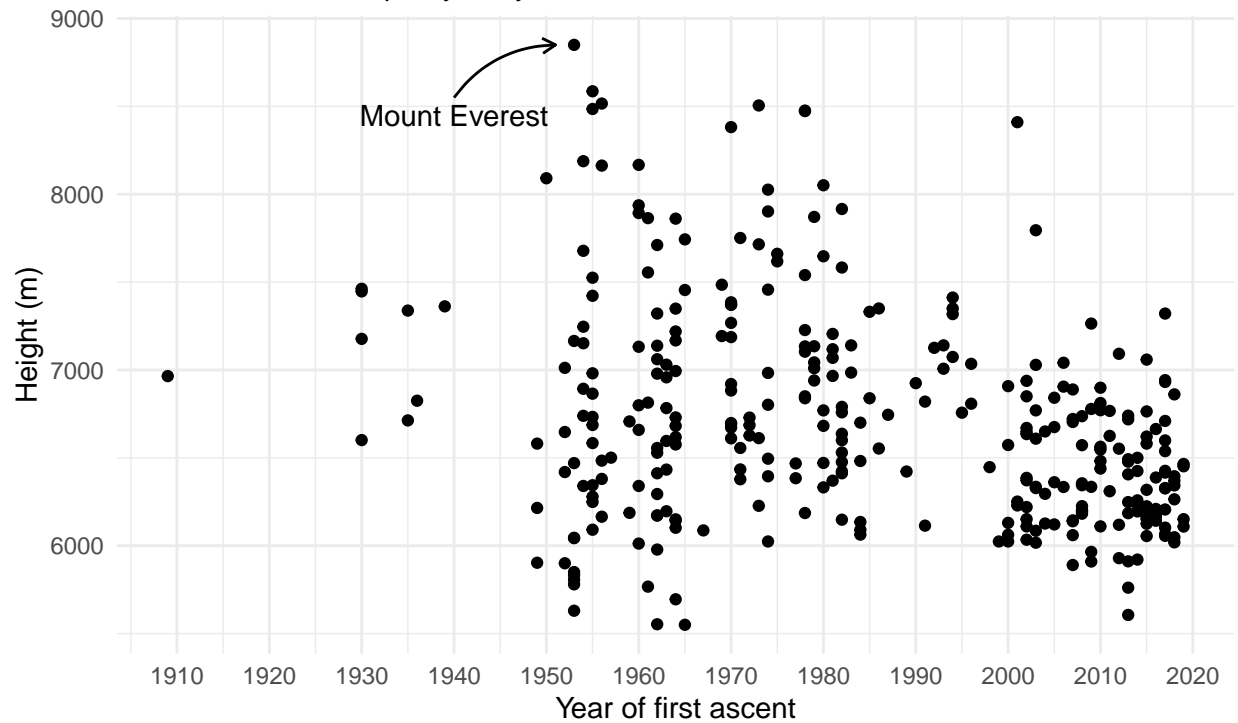


Source: The Himalayan Database

```
peaks %>%
  filter(first_ascent_year != 201 & !is.na(first_ascent_year)) %>%
  ggplot(aes(x = first_ascent_year, y = height_metres)) +
  geom_point() + theme_minimal() +
  scale_x_continuous(breaks = seq(1910, 2020, 10)) +
  labs(title = "Ofcourse !",
        subtitle = "We climbed Everest pretty early.",
        caption = "Source: The Himalayan Database",
        x = "Year of first ascent",
        y = "Height (m)") +
  annotate("text", 1940, 8450, label = "Mount Everest") +
  annotate(
    "curve",
    x = 1940,
    y = 8550,
    xend = 1951,
    yend = 8850,
    curvature = -0.25,
    arrow = arrow(length = unit(2, "mm"))
  )
```

Ofcourse !

We climbed Everest pretty early.



Source: The Himalayan Database

Questions

There are still some questions unanswered from the `peaks` dataset.

- Which country climbed most peaks first?
- Which peaks have multiple names?

Next Steps

- Analyse the members and expeditions data
- Find connections between different datasets
- Create a final analysis and data story

/TATA!/
/cube/