# 学一点数据分析-基本概念

第二节

# 数据分析的类别

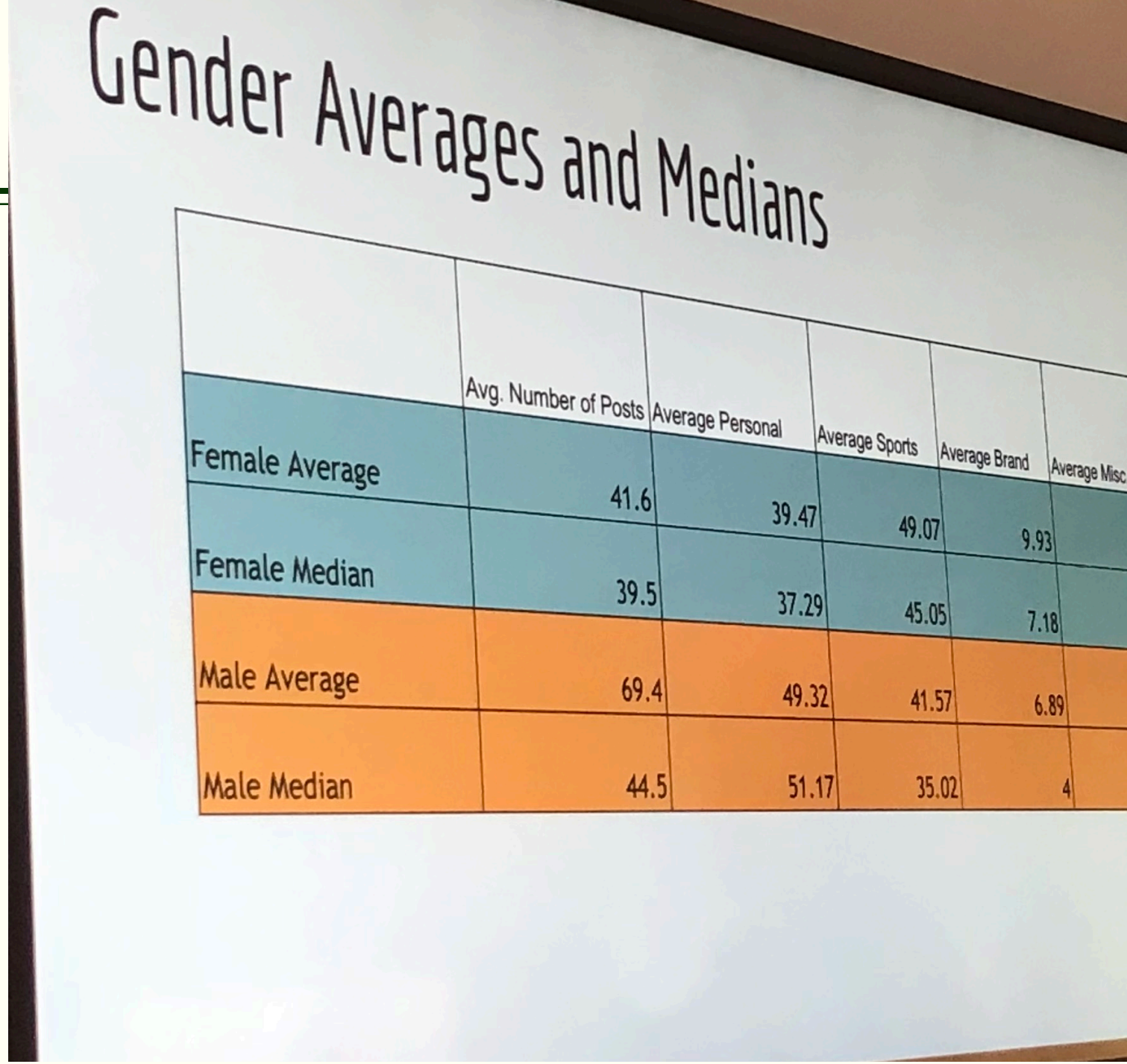
| | | |
|---|---|---|
| **数据分析** | **描述性数据分析** | **常用分析方法包括：<br>对比分析<br>平均分析<br>交叉分析** |
| | 探索性数据分析 | 常用分析方法包括：<br>相关分析<br>因子分析<br>回归分析 |
| | 验证性数据分析 | |

要掌握高级数据分析必须具有一定的统计学专业基础，如果媒体制作的数据新闻旨在做探索性数据分析或是验证性数据分析的话，则应该吸收专业人士参与报道团队。但是大多数数据新闻中的数据分析并不像想象的那么困难，它们多数属于基础的描述性统计分析，这种初级的数据分析是对一组数据的各种特征进行分析，以便于描述测量样本的各种特征及其所代表的总体的特征。要进入数据新闻领域，掌握初级数据分析是进行更为复杂的高级统计分析的基础。本节将着重介绍初级数据分析。

# Mistake in statistic

Sport report statistics
Female and male reporters

Posts on personal info
Posts on sports
Posts on Brand



# Revision on statistic

t-test anova
regression
logistic regression
chi-square

# 一、新闻中常见的统计学概念

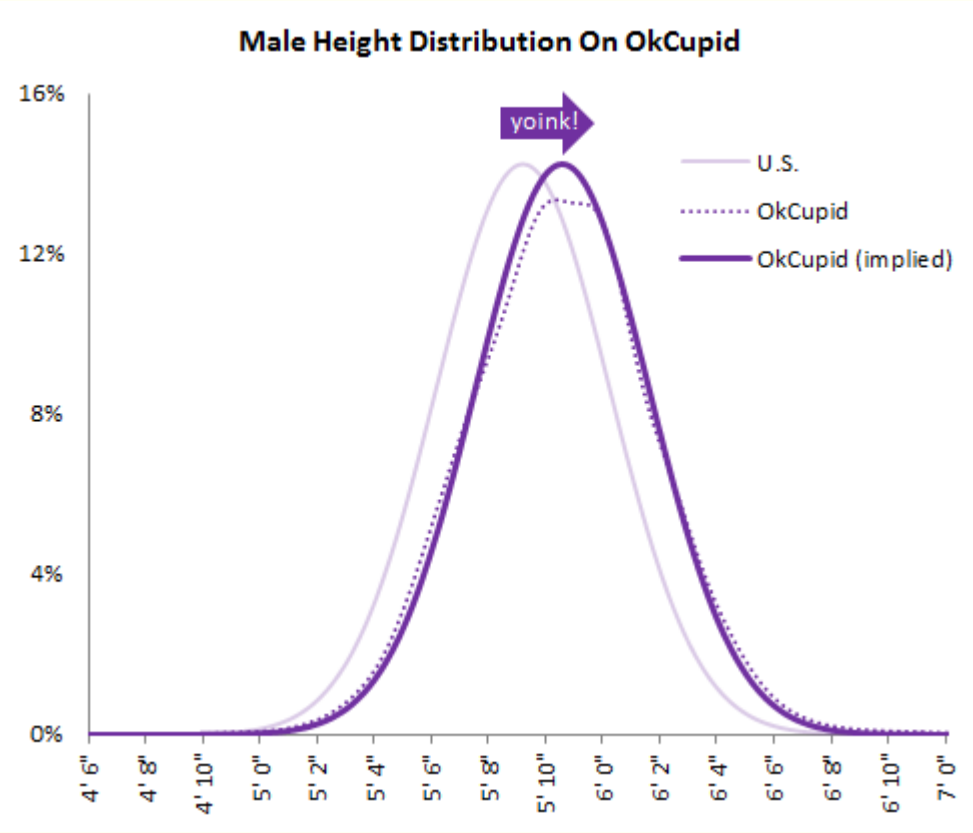
（一）描述统计数据的概念

1□均值
2□切尾均值
也叫截尾平均数，是指去掉观察值中的极端值后，根据剩下的观察值计算的平均数。
3□中位数
4□众数
5□百分比、百分点与百分比变化
百分比是相对数中的一种，所谓相对数，即表示一个数是另一个数的百分之几，也称为百分率或百分数。
百分点是一个很容易与百分比混淆的概念，它指不同时期以百分数的形式表示的相对指标的变化幅度，1个百分点=1%。
用新数值减去旧数值，所得的差再除以旧数值，得到了百分比变化。
6□人均数据
将要比较的数值总数除以人口总数（即基数），得到的是人均数据。
7□方差、标准差

# 案例：从正态分布图中发现网恋的身高谎言

虽然有关网恋的谎言经常见诸媒体，但是大多数媒体都是用采访个案的方式来报道的，而OkTrends作为美国最大的免费交友（恋人）网站OkCupid的博客，则通过大量数据分析来揭示人们网络交友时的一些现象。2010年7月7日，该博客发布了来自克里斯蒂·鲁德（Christian Rudder）的一项名为《有关网恋的大谎言》（The Big Lies People Tell in Online Dating）的调查性报道，揭示人们在交友网站填写资料档案时的一些普遍性的谎言。这项报道荣获了“信息之美”奖的数据新闻类大奖。

这则报道揭示的谎言分为四类，分别是身高的谎言、收入的谎言、近照的谎言和性别取向的谎言。其中在揭秘“身高的谎言”版块中，就使用了上文提及的正态分布指标。




https://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

# 一、新闻中常见的统计学概念

（二）常用的“指数”与“指标”

除了前述描述统计数据的概念之外，还有两个在报道中经常用到的词“指数”和“指标”也值得关注。

统计指数是对有关现象进行比较分析的相对比率。作为一种对比性的分析指标，“指数具有相对数的形式，通常表现为百分数。它表明，若把作为对比基准水平（基数）视为100，则所要考察的现象水平相当于基数的多少”
国内外常见的主要经济指数包括消费者价格指数（CPI）、零售价格指数、生产指数、生产者价格指数和股票价格指数。

为了读懂一些国民经济统计和社会发展统计公报，除了掌握上面这些“指数”之外，了解国民经济统计背后的“指标”体系也非常重要。如产品生产指标、收入分配指标、收入使用指标、投资积累指标、对外经济指标等。

# 1数据标准化案例

Exercise：做一个课堂练习：对以下指标体系提出你的改进意见

有一个众创空间的评估体系，其中两项三级指标是

单位面积产值 7分

单位面积孵化企业数 4分

评估方对其进行分段式打分处理，

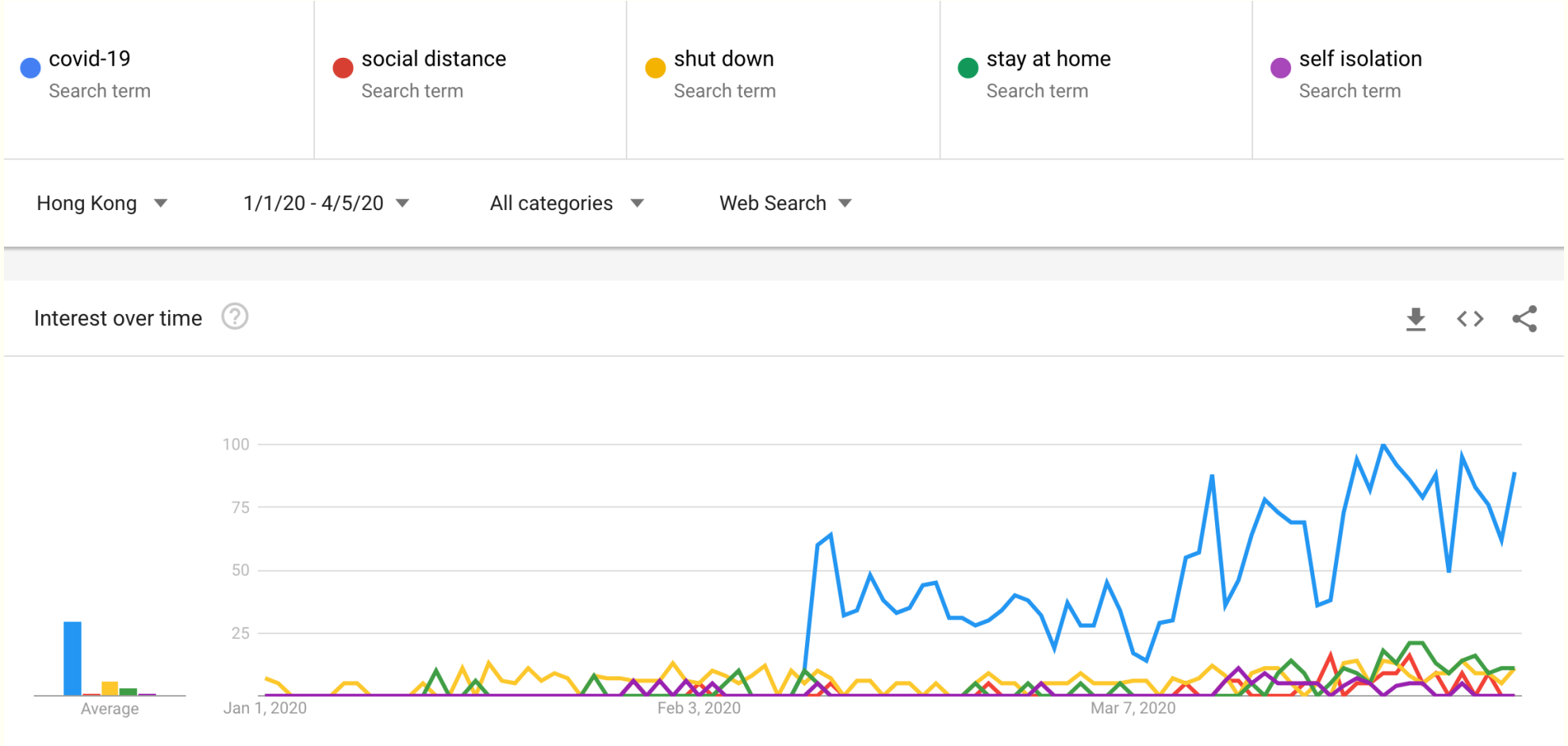
单位面积产值 7分（产值: 没有的0分，0到19的2分，19到31的4分，31到53的5分，高于53的7分）

单位面积孵化企业数 5分（孵化企业数: 没有的0分，0到17的1分，17到39的2分，39到57的3分，高于57的4分）

2数据转化比较分析案例

# Introduction of recent project related to the topic

How to tell the story about COVID-19Though different google search word?



covid-19
Search term

social distance
Search term

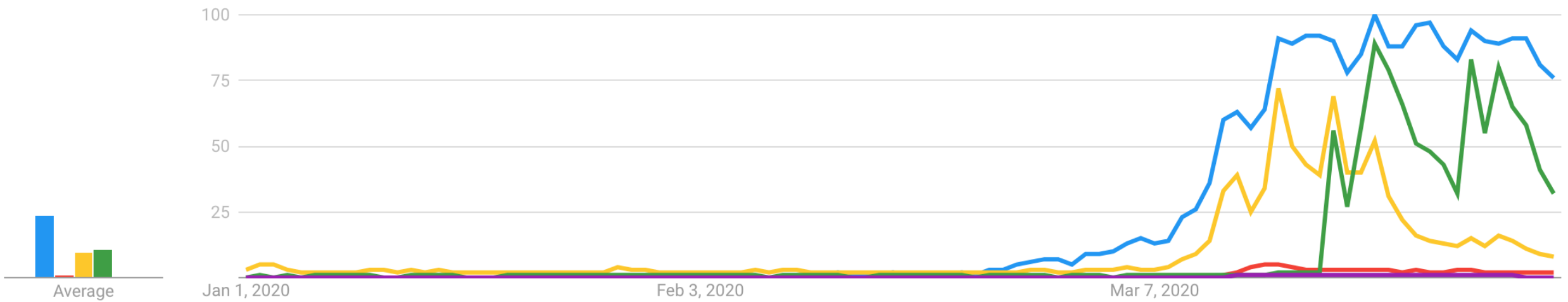
shut down
Search term

stay at home
Search term

self isolation
Search term

United States | 1/1/20 - 4/5/20 | All categories | Web Search

Interest over time



pneumonia
Search term

Shortness of breath
Search term

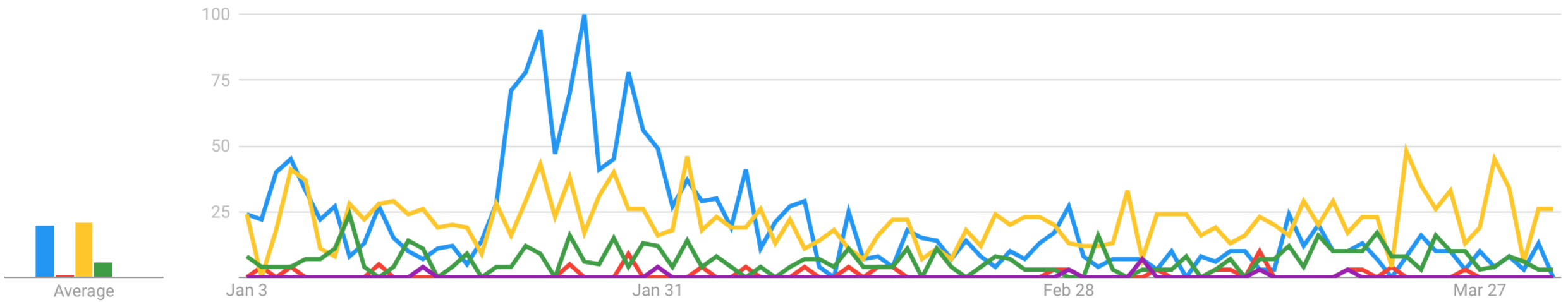
fever
Search term

cough
Search term

tiredness
Search term

Hong Kong | Past 90 days | All categories | Web Search

Interest over time



What are the challenges?

5 items?
Different states, different items?
Comparison but not exact number
Integer but not decimal
First 2 line?
<1

# Use python to do statistic analysis