HBS Survey Report

2020

A SURVEY PROJECT BY

VIJAYALAKSHMI RAMACHANDRAN,
FLORENCE PLACE,
AND ISHAN YASH

Objective

Few of my colleagues and I have gathered some data and have tried to implement the learnings of the HBS Business Analytics course on this raw dataset. We have tried to extrapolate the raw data by forming several Hypothesis and the objective is to find reliable representative of the given

THE PROCESS

- Data Collection.
- Data Processing.
- Statistical Analysis.
- Forming Hypothesis.
- To reject or accept the formed Hypothesis.

Data Collection

We have used Google forms to collect the data. We currently have 300 data points. We have collected data by asking the reader a certain set of questions. The Questions are:

- What is your nationality?
- You are from (Urban or Rural locality)
- How many children do you have?
- What is your child's age group?
- Imagine If your children would have to go to school in the current scenario (amidst a pandemic), how worried you would be? ["10" - extremely worried and "1" - not at all]

The aim was to form a dataset which could give us a glimpse of how would parents react to the whole schools' reopening scenario over the world.

Data Exploration and Preprocessing

After collecting the data we had to do some pre processing such as Data cleaning and Renaming the categories(The questions) to shorter names for ease.

E.g. Changing Indian, India, INDIA, Bhartiya to Indian and changing Null/NAN to zero.

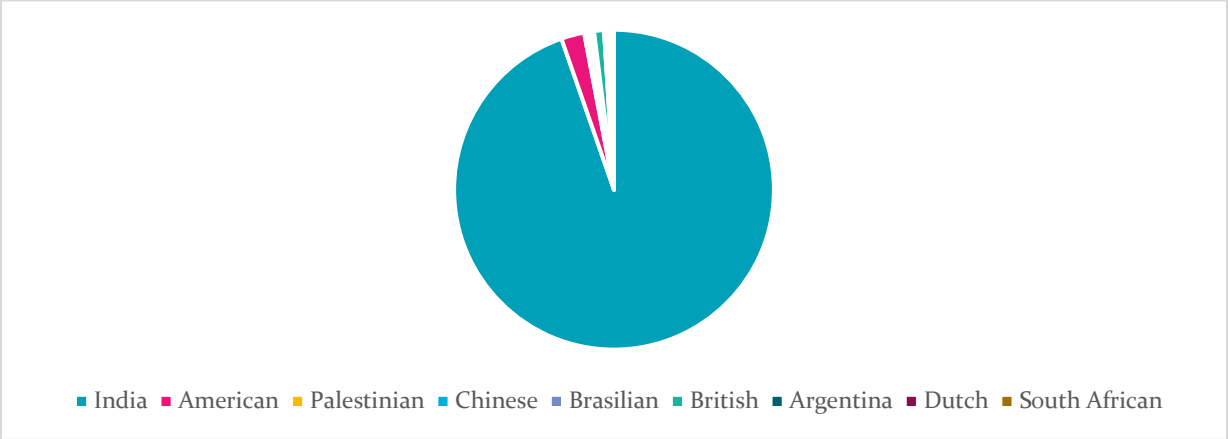


Figure 1. The split between nationalities in the data set

Statistical Analysis

The next step was to do some data analysis on the dataset, using Data Analysis column in excel (We have majorly used excel to perform Data Analysis and Cleaning in this project).

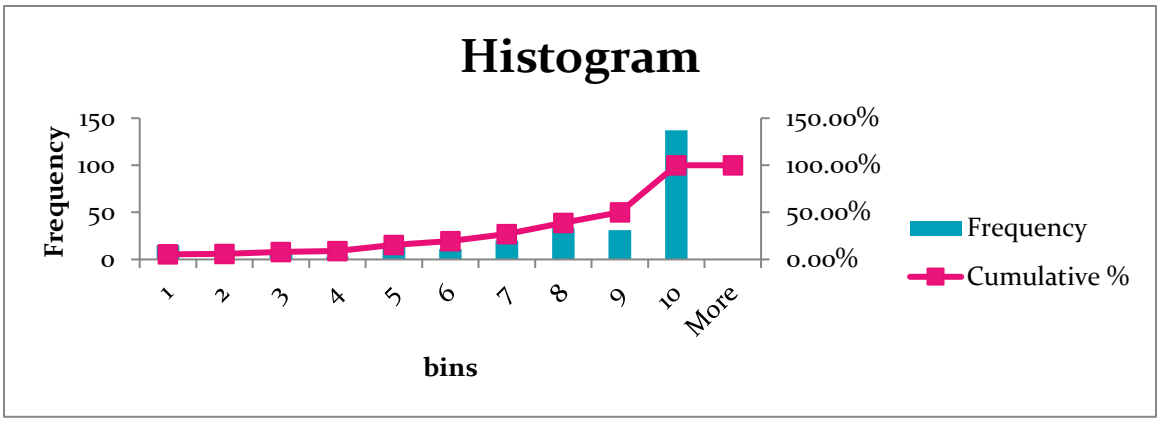
We have used Data Analysis column to perform:

1. Descriptive statistical Analysis
2. Form histogram(s)
3. To find confidence level (upper and lower bounds)
4. To perform T.Test (finding P value)

Forming Hypothesis

Part 1

Hypothesis 1: Variables are **no. of child** and the **worry factor**. Null Hypothesis is no difference in worry factor for parents having more than one child. Alt hypothesis would suggest that there will be a difference.



It is clear that most of the parents are worried but this can be used to extrapolate further information regarding whether this changes with the number of children they have or does the age group affect this in any way?

Therefore we can form a Hypothesis which has two variables i.e. no. of children and the worry factor.

The Null Hypothesis will be no difference in worry factor for parents having more than one child. Alt hypothesis would suggest that there will be a difference.

Hypothesis testing

So after the aforementioned steps we have come to check whether we are accepting the H_0 or rejecting it?

t-Test: Two-Sample Assuming Equal Variances

	no. of children	*hist. mean*
Mean	1.84	1.84
Variance	0.37	0.00
Observations	275.00	275.00
Pooled Variance	0.19	
Hypothesized Mean Difference	0.00	
df	548.00	
t Stat	-0.10	
P(T<=t) one-tail	0.46	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.92	
t Critical two-tail	1.96	

As we can see the p value is greater than 0.05 we can say we can not reject the null hypothesis

Therefore, there isn't a significance difference in the worry factor

Part 2

Worry factor in Rural vs Urban areas

H_0: There is no difference in worry factor between those in rural areas when compared to those in urban areas.
H_a: There is a difference in worry factor between those in rural areas when compared to those in urban areas.

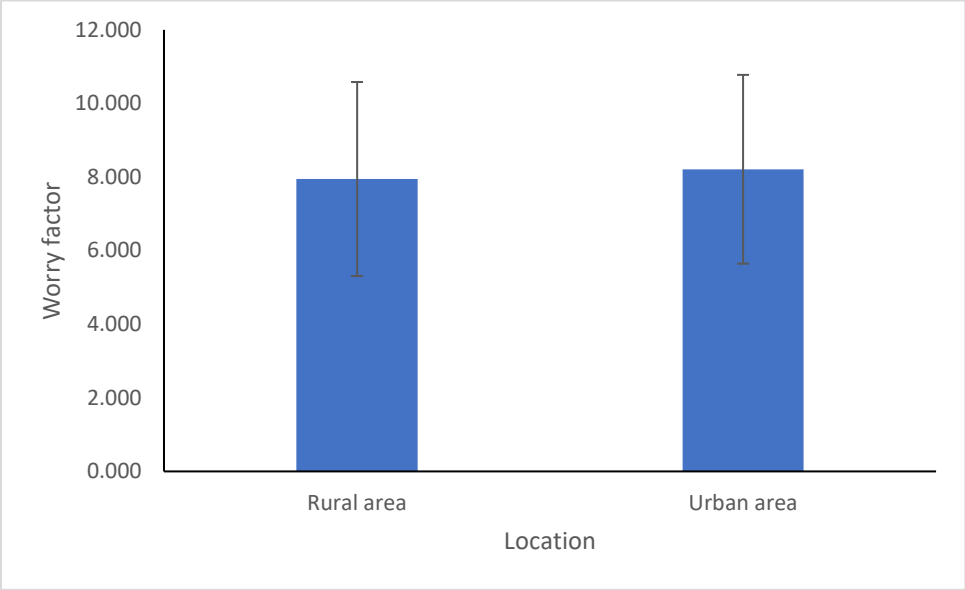


Figure 2. The difference in worry factor between rural and urban areas

Due to the similar means and fairly large standard deviations, shown in Figure 2, it is unlikely that there will be a significant difference in worry factor between the two areas.

Worry factor in different nationalities

H_0: There is no difference in worry factor between nationalities
H_a: There is a difference in worry factor between nationalities

Hypothesis testing

Worry factor in Rural vs Urban areas

t-Test: Two-Sample Assuming Equal Variances

	Rural area	*Urban area*
Mean	7.944	8.211
Variance	6.955	6.581
Observations	72.000	228.000
Pooled Variance	6.670	
Hypothesized Mean Difference	0.000	
df	298.000	
t Stat	-0.762	
P(T<=t) one-tail	0.223	
t Critical one-tail	1.650	
P(T<=t) two-tail	0.447	
t Critical two-tail	1.968	

We can be 95% confidence that there is no difference between the worry level of those in rural areas when compared to those in urban areas (p=0.223), therefore, we accept the null hypothesis.

Worry factor in different nationalities

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.164463637
R Square	0.027048288
Adjusted R Square	0.000300474
Standard Error	2.580443726
Observations	300

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	53.8679276	6.73349095	1.01123361	0.42750502
Residual	291	1937.67874	6.65868983		
Total	299	1991.54667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7.00	2.58	2.71	0.01	1.92	12.08
India	1.20	2.58	0.47	0.64	-3.88	6.29
American	0.86	2.76	0.31	0.76	-4.57	6.29
Palestinian	0.00	3.65	0.00	1.00	-7.18	7.18
Chinese	3.00	3.65	0.82	0.41	-4.18	10.18
Brasilian	2.00	3.65	0.55	0.58	-5.18	9.18
British	-1.67	2.98	-0.56	0.58	-7.53	4.20
Argentina	-1.00	3.65	-0.27	0.78	-8.18	6.18
Dutch	-3.00	3.65	-0.82	0.41	-10.18	4.18

We can be 95% confident there is no difference in levels of worry between nationalities ($p>0.05$), therefore, we accept the null hypothesis.