**METHODOLOGY**

We have an independent data yearly base in two groups 2019 and 2020 with insight attributes purchase expense, earnings and employees, so we applied different statistical tests T-test, Pearson Correlation and Cross Tabulation .It is used to compare a sample mean with a known population mean or some other meaningful, fixed value.

**Data Formatting**

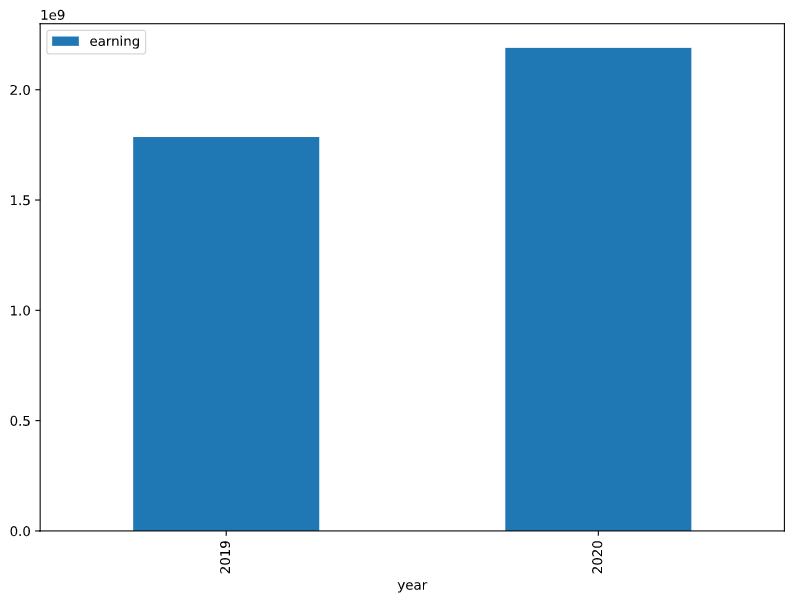
Data formatting was crucial as data was raw, and data was distinct at some point because data gathered from distributed databases. To relate the data and make it meaningful year and month columns were merged in the date column to get a sense of every month earning data. In current\_working\_status column 1 taking as business is open and working with us and 0 as closed.

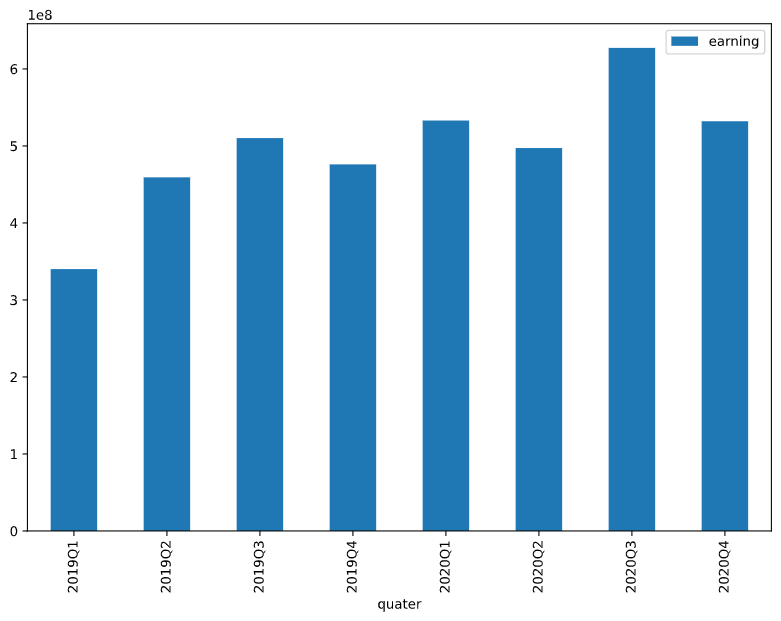
**Data Cleaning**

Some fields have null values. Missing values were replaced by the mostly occurring value of that attribute subject to its related output, which means that we picked all the example points which have the same output to that example point, and we replace the missing value with mostly occurring values of the attribute. Remove the store id attribute because it's unique and unusable.

1. **Revenue comparison during peak and off-peak time covid**

We have monthly data from january 2019 to December 2020. During this time period in the first half of 2020 COVID is at its peak and people have more anxiety and psychological problems during this time. So we have a hypothesis on this basis that during 2020 more repairing jobs will generate more revenue. In fig 1 we can see that in 2019 revenue is 1.785576 E09 and in 2020 it is 2.189717 E+09.

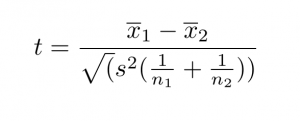




To support our hypothesis we perform a T-test.

### **T-test**

The t-test is a [**parametric** test](https://www.scribbr.com/statistics/statistical-tests/#parametric) of difference, meaning that it makes the same assumptions about your data as other parametric tests. The formula for the two-sample t-test (a.k.a. the Student’s t-test) is shown below.



In this formula, ***t*** is the t-value, ***x*1** and ***x*2** are the means of the two groups being compared, ***s*2** is the pooled standard error of the two groups, and ***n*1** and ***n*2** are the number of observations in each of the groups. A larger *t*-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.

We can compare your calculated *t*-value against the values in a critical value chart to determine whether your *t*-value is greater than what would be expected by chance. If so, you can reject the null hypothesis and conclude that the two groups are in fact different.

**Independent sample T test between year and earnings**

Ho = The earnings in 2020 is greater than or equal to earnings in 2019

H1 = The earnings in 2020 is less than the earnings in 2019

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | year | N | Mean | Std. Deviation | Std. Error Mean |
| Earning | 2019 | 5352 | 270776.3249 | 3254492.39758 | 44486.18965 |
| 2020 | 5352 | 288565.0915 | 3467708.82066 | 47400.67986 |

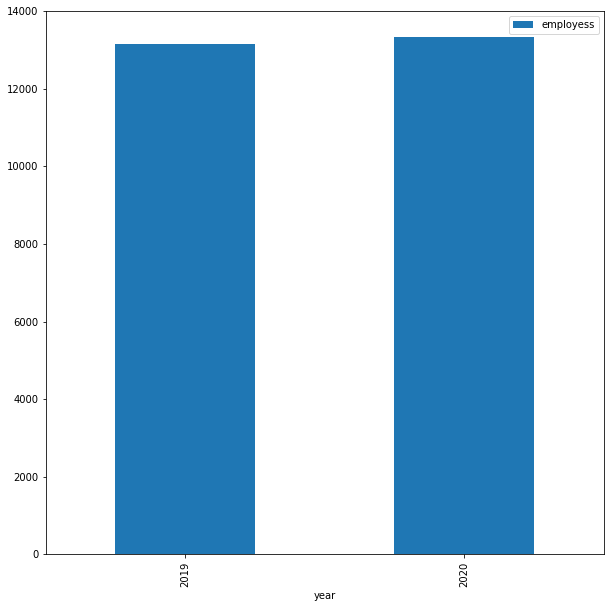
|  | |  | | | | | | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |  |
| Lower | Upper |  |
| earning | Equal variances assumed | -.274 | 10702 | .784 | -17788.76657 | 65006.50369 | -145213.58390 | 109636.05076 |  |
| Equal variances not assumed | -.274 | 10659.192 | .784 | -17788.76657 | 65006.50369 | -145213.64179 | 109636.10864 |  |

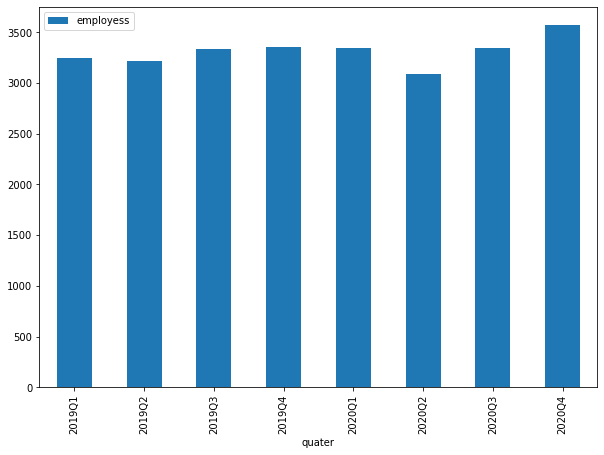
An independent samples t-test was used to compare the earnings of 2019 and 2020. According to group statistics the number of earnings in 2019(N=5352) and in 2020(N=5352). The t-test was statistically significant, with mean score. In table it can be observed that, p>.05, Therefore, the null hypothesis is failed to reject. It can be concluded that the earnings in 2020 were greater than or equal the earnings in 2019.

1. **workforce comparison during peak and off-peak time covid**

Another effect of covid on all over the world can see the job loss in a number of businesses mostly which includes physical work. But on the contrary there are some businesses which boom during this time. To notice this effect in our business some comparisons are performed. As we have a hypothesis during covid anxiety and depression increase among people thats we see in our earning its increase during 2020 so for now we have a hypothesis that in 2020 we will have more employees then 2019. As it is visible in chart fig 2 during 2019 stores have 13152 employees and in 2020 these stores have 13341. Which is not much difference but means a lot as we noticed business downsized.

To support our hypothesis we perform a T-test.





**(T-test info if we want to add)**

As we have an independent attribute of time period and employess.

**Independent sample T test between year and no of employees**

Ho = The no of employees in 2020 were less than or equal to no of employees in 2019.

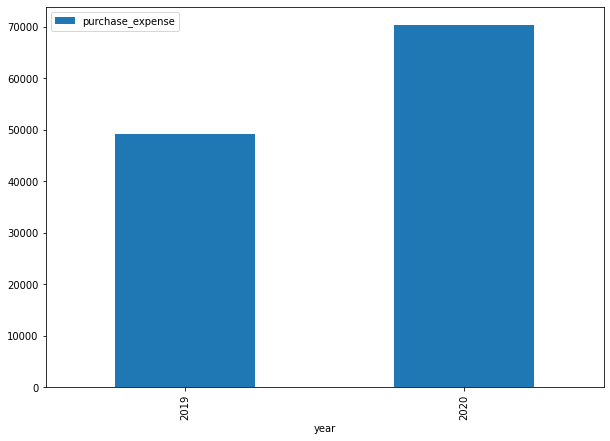
H1 = The no of employees in 2020 were greater than the no of employees in 2019.

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | year | N | Mean | Std. Deviation | Std. Error Mean |
| employess | 2019 | 5352 | 2.46 | 1.978 | .027 |
| 2020 | 5352 | 2.49 | 2.015 | .028 |

|  | |  | | | | | | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |  |
| Lower | Upper |  |
| employess | Equal variances assumed | -.915 | 10702 | .360 | -.035 | .039 | -.111 | .040 |  |
| Equal variances not assumed | -.915 | 10698.376 | .360 | -.035 | .039 | -.111 | .040 |  |

An independent samples t-test was used to compare the no of employees of 2019 and 2020. According to group statistics the number of no of employees in 2019(N=5352) and in 2020(N=5352). The t-test was statistically significant, with mean score. In table it can be observed that, p>.05, Therefore, the null hypothesis is failed to reject. It can be concluded that the no of employees in 2020 were less than or equal to the employees in 2019.

1. **repair industry expenses comparison during peak and off-peak time covid**

****

**(Need to add some intro info about purchase expense and its need)**

**Independent sample T test between year and purchase expense**

Ho = The purchase expense in 2020 is greater than and equal to purchase expense in 2019

H1 = The purchase expense in 2020 is less than purchase expense in 2019

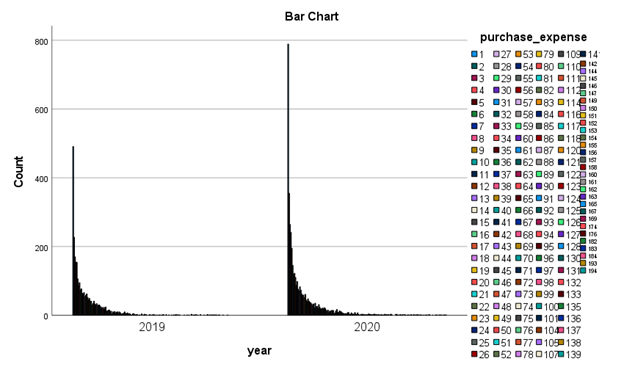
| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | year | N | Mean | Std. Deviation | Std. Error Mean |
| purchase expense | 2019 | 2982 | 16.52 | 23.538 | .431 |
| 2020 | 4169 | 16.87 | 26.559 | .411 |

|  | |  | | | | | | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |  |
| Lower | Upper |  |
| purchase expense | Equal variances assumed | -.575 | 7149 | .565 | -.349 | .608 | -1.541 | .842 |  |
| Equal variances not assumed | -.586 | 6830.944 | .558 | -.349 | .596 | -1.517 | .819 |  |

An independent samples t-test was used to compare the purchase expense of 2019 and 2020. According to group statistics the number of purchases in 2019(N=2982) and in 2020(N=4169). The t-test was statistically significant, with mean score of 2019 (M=16.52, SD=23.538) & for 2020 (M=16.87, SD=26.59). In table it can be observed that, p>.05. Therefore, the null hypothesis is failed to reject. It can be concluded that the purchase expense in 2020 were greater than the purchase expense in 2019.

**Cross Tabulation performed in between years effects on purchase expense.**

The cross tabulation of year and purchase expense can be observed in bar chart for the comparison of the number of purchases each year.



| Attribute name | Data type | Description |
| --- | --- | --- |
| Year | int64 | This attribute represent the year of data , and range between 2019 and 2020 |
| Month | int64 | This attribute represent the months of data |
| Employees | int64 | This attribute represent no’s employees of business during every month |
| Earning | int64 | This attribute represent total no’s of sales during every month |
| Date | string | This attribute is combination of year and month |
| Store\_id | int64 | This attribute represents business id as reference |
| Country | string | This attribute represents business belong to country |
| current\_working\_status | int64 | This attribute represents business status is active or not |
| Purchases\_expense | int64 | This attribute represent total no’s of expenses during every month |

dtype: object

\frac{\sum (x\_i - \overline{x})(y\_i - \overline{y})}{\sqrt{\sum{(x\_i - \overline{x})ˆ2}}{\sqrt{\sum{(y\_i - \overline{y})ˆ2}}}}

**Result & Conclusion**

| **Yearly Comparison Table** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Sr.No** | **Year** | **Employees** | **Purchase Expense** | **Revenue** | **Result** |
| 1 | 2019 | 13152 | 49268 | 1449192626 | After Comparing Yearly data we conclude Covid-19 is a loss year in each term as compared to 2020. |
| 2 | 2020 | 13341 | 70336 | 1544398168 |

| **Complete Dataset** | |
| --- | --- |
| **Attributes** | **Total Values** |
| Country | 37 |
| Business | 446 |
| Employees | 26493 |
| Records | 10704 |
| Revenue | 2993590794 |
| Purchase Expense | 119604 |

| **TOP 5 COUNTRY COMPARISON** | | |
| --- | --- | --- |
| **Attributes** | **2019** | **2020** |
| Employee | 10683 | 10996 |
| Purchase Expense | 38956 | 56472 |
| Revenue | 1357546000 | 1442299000 |