Exploratory Data Analysis on the Titanic Dataset

Submitted by: *Lavanya Gupta*
**Internship Project*
Date: *July-August 2025*

1. Introduction

The Titanic dataset is a widely used dataset for practicing data cleaning, exploratory data analysis, and predictive modeling tasks. It contains information about the passengers aboard the Titanic, such as age, sex, class, fare, and survival status.

In this project, I have performed **data cleaning** and **exploratory data analysis (EDA)** using **Google Sheets**. The objective is to identify patterns in the data related to **survival**, including variations in survival rates by gender, social class, age, and fare.

2. Dataset Description

The dataset used is `train.csv` from Kaggle's Titanic Machine Learning Competition. It includes 891 rows and 12 columns. Key columns used in the analysis are:

- `Survived`: Whether the passenger survived (1) or not (0)
- `Pclass`: Passenger class (1st, 2nd, 3rd)
- `Sex`: Gender of the passenger
- `Age`: Age of the passenger
- `SibSp`, `Parch`: Number of siblings/spouses or parents/children aboard
- `Fare`: Fare paid by the passenger
- `Embarked`: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3. Data Cleaning

The following data-cleaning steps were performed:

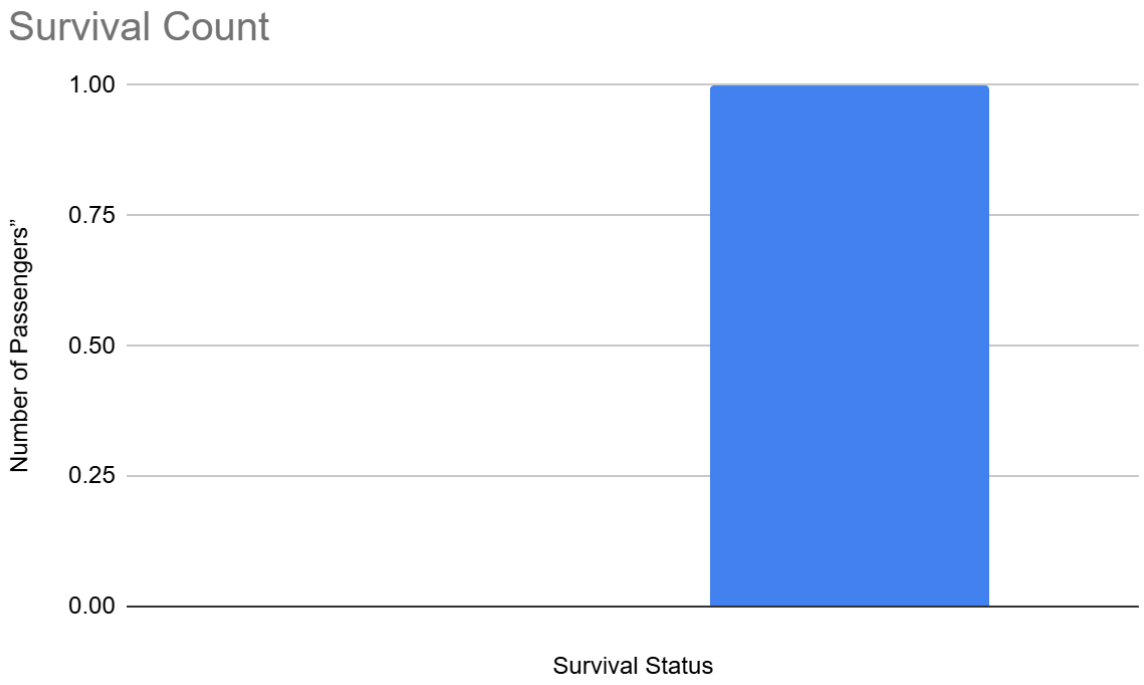
- **Filled missing `Age` values** using the median age.
- **Verified that the `Embarked` column** had no missing values.

- **Removed `Cabin` column** due to too many missing entries.
- **Created a new column `Age Filled`** to use for analysis.

4. Exploratory Data Analysis (EDA)

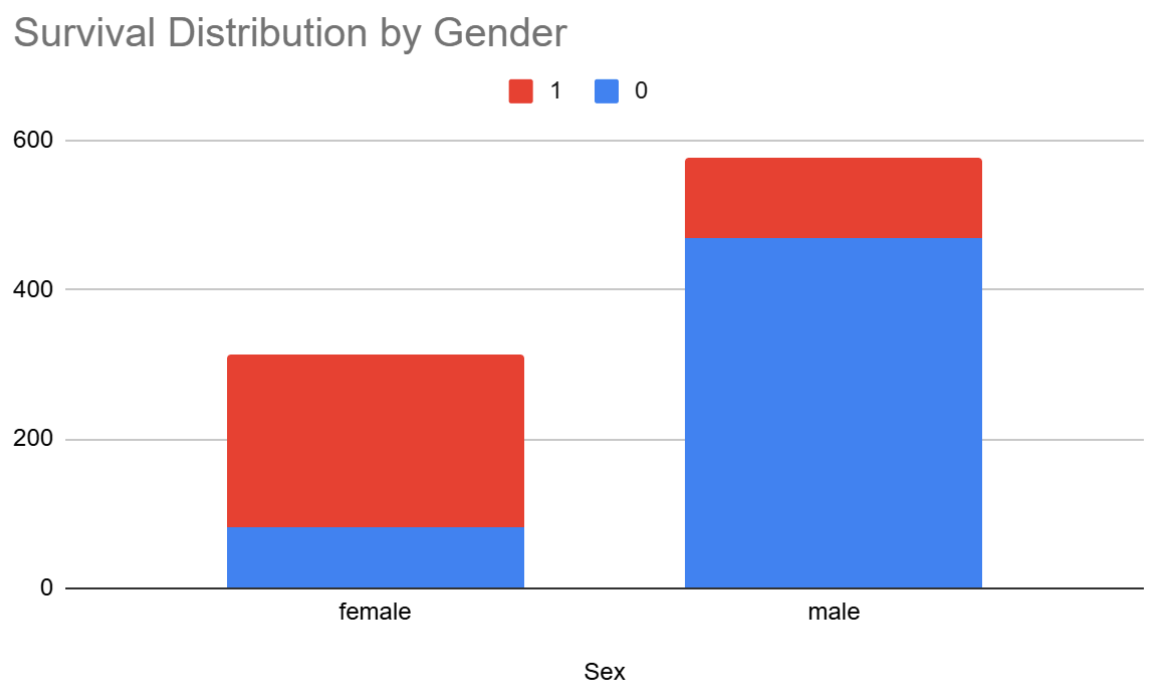
4.1 Survival Count

Chart: Survival Count (0 = Did not survive, 1 = Survived)



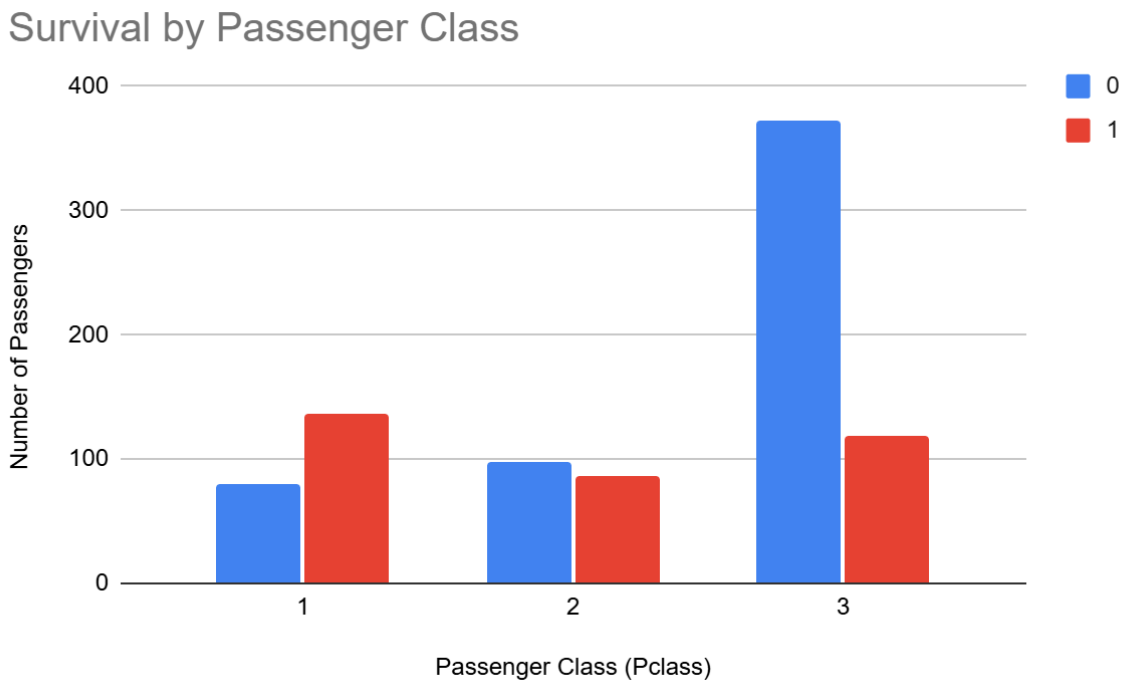
Out of 891 passengers, around 38% survived and 62% did not.

4.2 Survival by Gender



The survival rate was significantly higher for **females (about 74%)** than for males (about 19%). This supports the "women and children first" evacuation strategy.

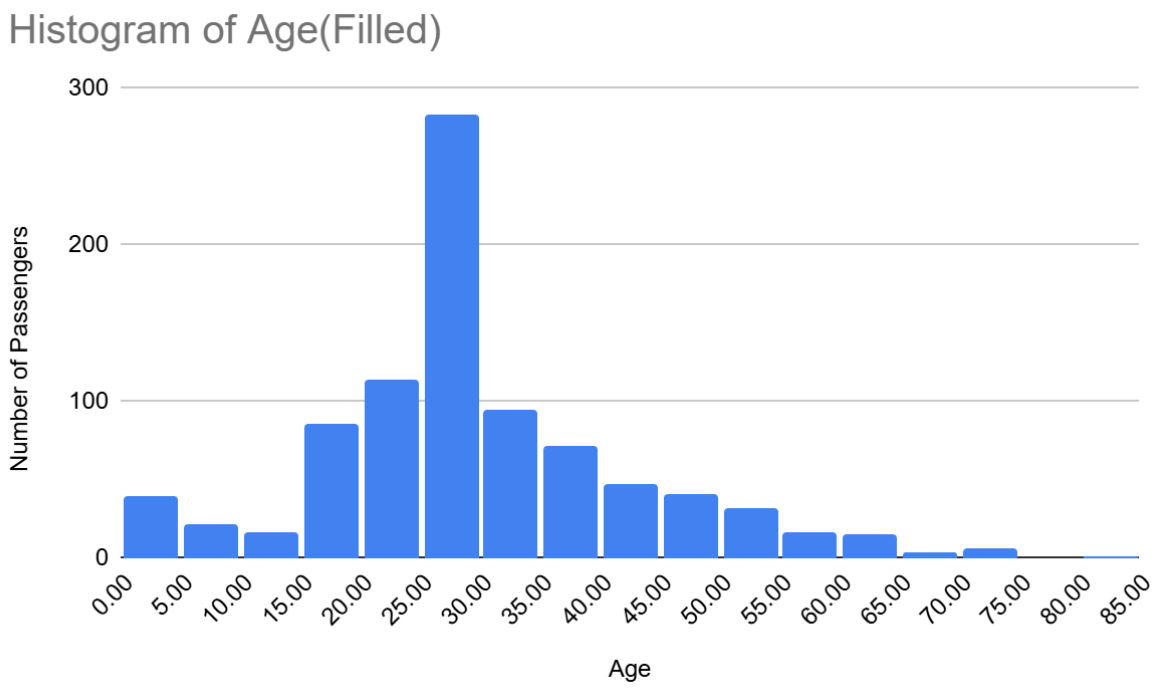
4.3 Survival by Passenger Class



Passengers in **1st class** had a much higher chance of survival compared to those in 3rd class.

4.4 Age Distribution

Chart: Age Distribution



Most passengers were between **20 and 40 years old**. A smaller number were children under 10 or adults over 60.

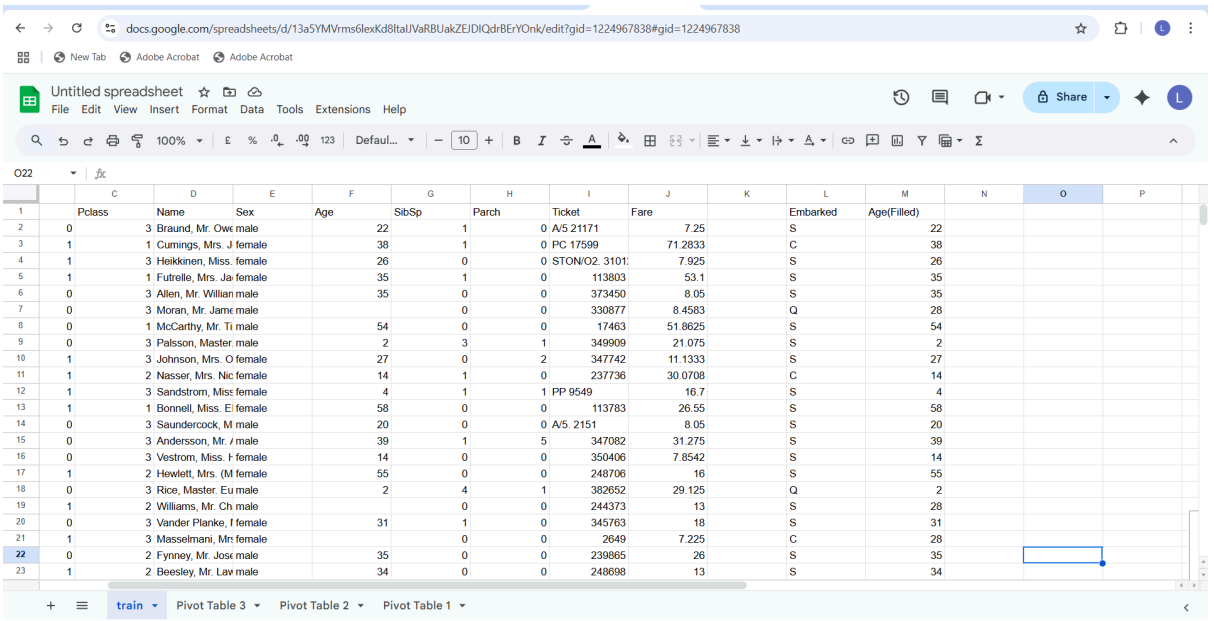
4. Conclusion

The analysis of the Titanic dataset revealed clear patterns in survival rates:

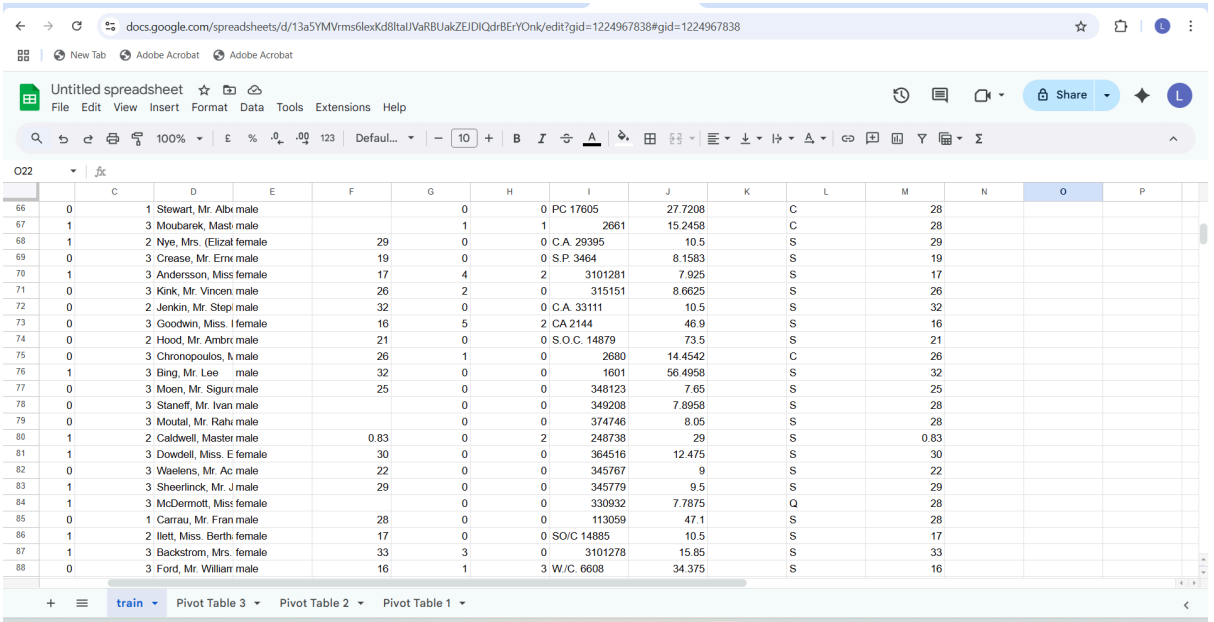
- **Females** had a significantly higher chance of survival.
- **1st class passengers** were more likely to survive than 2nd and 3rd class.
- **Age** played a moderate role, with some survival preference for younger passengers.

This project provided hands-on experience with **data cleaning**, **pivot tables**, **charts**, and generating actionable **insights** from real-world data using **Google Sheets**.

5. Screenshot of cleaned data

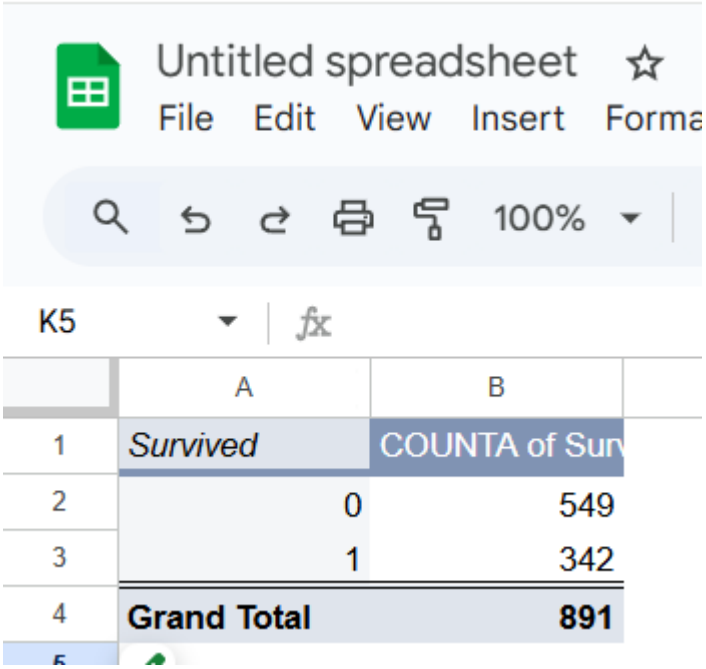


	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Age(Filled)
0	3	Braund, Mr. Owe	male	22	1	0	A/5 21171	7.25	S	22
1	1	Cumings, Mrs. J	female	38	1	0	PC 17599	71.2833	C	38
1	3	Heikkinen, Miss.	female	26	0	0	STON/O2. 3101	7.925	S	26
1	1	Futrelle, Mrs. Ja	female	35	1	0	113803	53.1	S	35
0	3	Allen, Mr. Willian	male	35	0	0	373450	8.05	S	35
0	3	Moran, Mr. Jame	male		0	0	330877	8.4583	Q	28
0	1	McCarthy, Mr. Ti	male	54	0	0	17463	51.8625	S	54
0	3	Palsson, Master.	male	2	3	1	349909	21.075	S	2
1	3	Johnson, Mrs. O	female	27	0	2	347742	11.1333	S	27
1	2	Nasser, Mrs. Nic	female	14	1	0	237736	30.0708	C	14
1	3	Sandstrom, Miss	female	4	1	1	PP 9549	16.7	S	4
1	1	Bonnell, Miss. El	female	58	0	0	113783	26.55	S	58
0	3	Saundercock, M	male	20	0	0	A/5. 2151	8.05	S	20
0	3	Andersson, Mr. A	male	39	1	5	347082	31.275	S	39
0	3	Vestrom, Miss. H	female	14	0	0	350406	7.8542	S	14
1	2	Hewlett, Mrs. (M	female	55	0	0	248706	16	S	55
0	3	Rice, Master. Eu	male	2	4	1	382652	29.125	Q	2
1	2	Williams, Mr. Ch	male		0	0	244373	13	S	28
0	3	Vander Planke, I	female	31	1	0	345763	18	S	31
1	3	Masselmani, Mrs	female		0	0	2649	7.225	C	28
0	2	Fynney, Mr. Jose	male	35	0	0	239865	26	S	35
1	2	Beesley, Mr. Law	male	34	0	0	248698	13	S	34

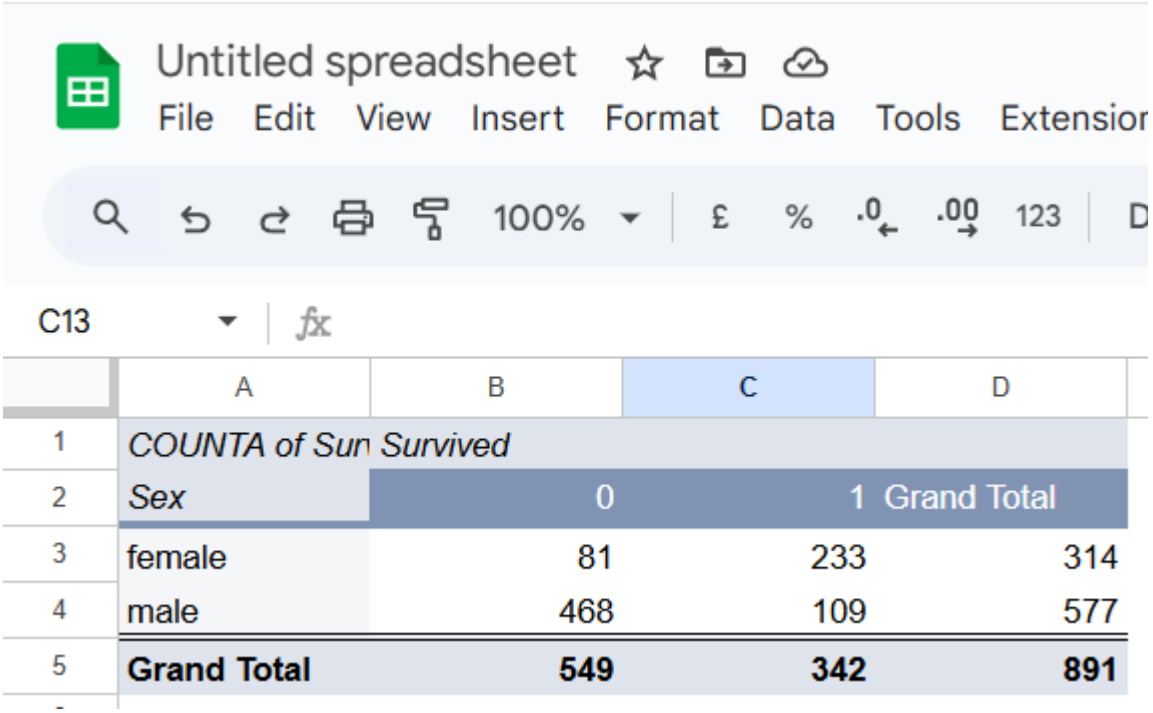


0	1	Stewart, Mr. Albe	male		0	0	PC 17605	27.7208	C	28
1	3	Moubarek, Maste	male		1	1	2661	15.2458	C	28
1	2	Nye, Mrs. (Elizat	female	29	0	0	C.A. 29395	10.5	S	29
0	3	Crease, Mr. Ernes	male	19	0	0	S.P. 3464	8.1583	S	19
1	3	Andersson, Miss	female	17	4	2	3101281	7.925	S	17
0	3	Kink, Mr. Vincen	male	26	2	0	315151	8.6625	S	26
0	2	Jenkin, Mr. Step	male	32	0	0	C.A. 33111	10.5	S	32
0	3	Goodwin, Miss. L	female	16	5	2	CA 2144	46.9	S	16
0	2	Hood, Mr. Ambr	male	21	0	0	S.O.C. 14879	73.5	S	21
0	3	Chronopoulos, M	male	26	1	0	2680	14.4542	C	26
1	3	Bing, Mr. Lee	male	32	0	0	1601	56.4958	S	32
0	3	Moen, Mr. Sigurd	male	25	0	0	348123	7.65	S	25
0	3	Staneff, Mr. Ivan	male		0	0	349208	7.8958	S	28
0	3	Moutal, Mr. Rahi	male		0	0	374746	8.05	S	28
1	2	Caldwell, Master	male	0.83	0	2	248738	29	S	0.83
1	3	Dowdell, Miss. E	female	30	0	0	364516	12.475	S	30
0	3	Waelens, Mr. Ac	male	22	0	0	345767	9	S	22
1	3	Sheerlinck, Mr. J	male	29	0	0	345779	9.5	S	29
1	3	McDermott, Miss	female		0	0	330932	7.7875	Q	28
0	1	Carrau, Mr. Fran	male	28	0	0	113059	47.1	S	28
1	2	Ilett, Miss. Berth	female	17	0	0	SO/C 14885	10.5	S	17
1	3	Backstrom, Mrs.	female	33	3	0	3101278	15.85	S	33
0	3	Ford, Mr. Willian	male	16	1	3	W./C. 6608	34.375	S	16

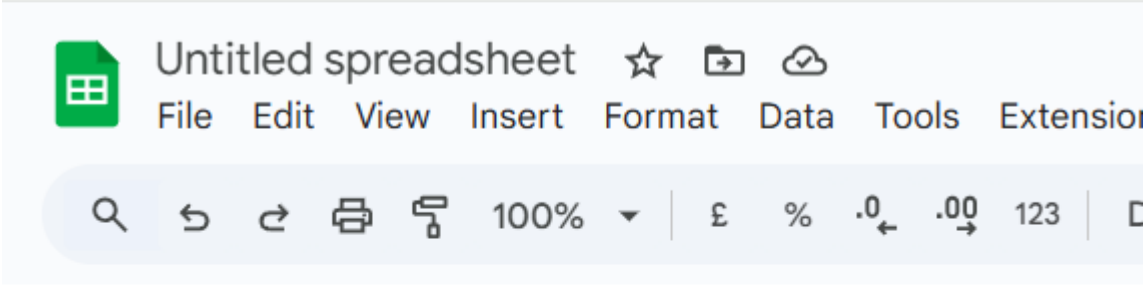
6. Raw Pivot Tables(before charting)



	A	B
1	*Survived*	COUNTA of Sur
2	0	549
3	1	342
4	**Grand Total**	**891**



	A	B	C	D
1	*COUNTA of Sur*	*Survived*		
2	*Sex*	0	1	Grand Total
3	female	81	233	314
4	male	468	109	577
5	**Grand Total**	**549**	**342**	**891**



J17

	A	B	C	D
2	*Pclass*	0	1	Grand Total
3	1	80	136	216
4	2	97	87	184
5	3	372	119	491
6	**Grand Total**	**549**	**342**	**891**