

```
=====
TITANIC SURVIVAL PREDICTION - CSYE 7200 ASSIGNMENT
=====
```

```
[STEP 1] Loading Datasets...
```

```
Train Dataset: 891 rows, 12 columns
```

```
Test Dataset: 418 rows, 11 columns
```

```
Train Dataset Schema:
```

```
root
```

```
  |-- PassengerId: integer (nullable = true)
  |-- Survived: integer (nullable = true)
  |-- Pclass: integer (nullable = true)
  |-- Name: string (nullable = true)
  |-- Sex: string (nullable = true)
  |-- Age: double (nullable = true)
  |-- SibSp: integer (nullable = true)
  |-- Parch: integer (nullable = true)
  |-- Ticket: string (nullable = true)
  |-- Fare: double (nullable = true)
  |-- Cabin: string (nullable = true)
  |-- Embarked: string (nullable = true)
```

```
First 5 rows:
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	NULL	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	NULL	S

only showing top 5 rows

```
=====
STEP 2: EXPLORATORY DATA ANALYSIS (EDA)
=====
```

Following up on Assignment 1 with additional insights

---

```
-----  
ASSIGNMENT 1 FOLLOW-UP ANALYSIS  
-----
```

```
[Assignment 1 Q1] AVERAGE TICKET FARE BY CLASS
```

---

Pclass	Avg_Fare	Min_Fare	Max_Fare	StdDev_Fare
1	84.15468749999992	0.0	512.3292	78.38037264672882
2	20.66218315217391	0.0	73.5	13.417398756149339
3	13.675550101832997	0.0	69.55	11.778141704387311

[Assignment 1 Q2] SURVIVAL PERCENTAGE BY CLASS

---

Pclass	Total	Survived	Survival_Percentage
1	216	136	62.96
2	184	87	47.28
3	491	119	24.24

Highest survival rate: Class 1 with 62.96%

[Assignment 1 Q3] PASSENGERS MATCHING ROSE DEWITT BUKATER

---

Rose: 17 years old, female, 1st class, traveling with 1 parent (mother)

Number of passengers matching Rose: 0

PassengerId	Name	Age	Sex	Pclass	Parch	Survived

[Assignment 1 Q4] PASSENGERS MATCHING JACK DAWSON

-----  
Jack: 19-20 years old, male, 3rd class, no relatives onboard

Number of passengers matching Jack: 23

PassengerId	Name	Age	Sex	Pclass	SibSp	Parch	Survived
13	Saundercock, Mr. William Henry	20.0	male	3	0	0	0
68	Crease, Mr. Ernest James	19.0	male	3	0	0	0
92	Andreasson, Mr. Paul Edvin	20.0	male	3	0	0	0
132	Coelho, Mr. Domingos Fernando	20.0	male	3	0	0	0
144	Burke, Mr. Jeremiah	19.0	male	3	0	0	0
284	Dorking, Mr. Edward Arthur	19.0	male	3	0	0	1
303	Johnson, Mr. William Cahoone Jr	19.0	male	3	0	0	0
373	Beavan, Mr. William Thomas	19.0	male	3	0	0	0
379	Betros, Mr. Tannous	20.0	male	3	0	0	0
380	Gustafsson, Mr. Karl Gideon	19.0	male	3	0	0	0
442	Hampe, Mr. Leon	20.0	male	3	0	0	0
567	Stoytcheff, Mr. Ilia	19.0	male	3	0	0	0
576	Patchett, Mr. George	19.0	male	3	0	0	0
641	Jensen, Mr. Hans Peder	20.0	male	3	0	0	0
647	Cor, Mr. Liudevit	19.0	male	3	0	0	0
683	Olsvigen, Mr. Thor Anderson	20.0	male	3	0	0	0
688	Dakic, Mr. Branko	19.0	male	3	0	0	0
716	Soholt, Mr. Peter Andreas Lauritz Andersen	19.0	male	3	0	0	0
726	Oreskovic, Mr. Luka	20.0	male	3	0	0	0
763	Barah, Mr. Hanna Assi	20.0	male	3	0	0	1

only showing top 20 rows

only showing top 20 rows

[Assignment 1 Q5] AGE GROUPS (10-YEAR INTERVALS) VS FARE & SURVIVAL

AgeGroup	Count	Avg_Fare	Survival_Rate	Survival_Percentage
1-10	62	30.57667903225807	0.6129032258064516	61.29
11-20	102	32.53513235294118	0.4019607843137255	40.2
21-30	220	27.278937272727294	0.35	35.0
31-40	167	40.377294011976055	0.437125748502994	43.71
41-50	89	38.00229662921349	0.38202247191011235	38.2
51-60	48	47.93333333333334	0.4166666666666667	41.67
61-70	19	48.36754210526315	0.3157894736842105	31.58
71-80	6	30.19723333333333	0.0	0.0
81-90	1	30.0	1.0	100.0

Age group most likely to survive: 81-90 with 100.0% survival rate

-----  
ADDITIONAL EXPLORATORY ANALYSIS  
-----

## ADDITIONAL EXPLORATORY ANALYSIS

### [6] OVERALL SURVIVAL RATE

Total: 891, Survived: 342 (38.38%)

Survived	count
1	342
0	549

### [7] SURVIVAL BY GENDER

Sex	Total	Survived	Rate	Survival_Pct
female	314	233	0.7420382165605095	74.2
male	577	109	0.18890814558058924	18.89

↑

↓

## [8] MISSING VALUES ANALYSIS

⤵

PassengerId : 0 (0.00%)

⤶

Survived : 0 (0.00%)

⤷

Pclass : 0 (0.00%)

⤸

Name : 0 (0.00%)

Sex : 0 (0.00%)

Age : 177 (19.87%)

SibSp : 0 (0.00%)

Parch : 0 (0.00%)

Ticket : 0 (0.00%)

Fare : 0 (0.00%)

Cabin : 687 (77.10%)

Embarked : 2 (0.22%)

## [9] FAMILY SIZE ANALYSIS

FamilySize	Count	Rate	Survival_Pct
1	537	0.30353817504655495	30.35
2	161	0.5527950310559007	55.28
3	102	0.5784313725490197	57.84
4	29	0.7241379310344828	72.41
5	15	0.2	20.0
6	22	0.13636363636363635	13.64
7	12	0.3333333333333333	33.33
8	6	0.0	0.0
11	7	0.0	0.0

▽

↔

E↓

🖨️

⌫

[10] EMBARKED PORT STATISTICS

Embarked	Total	Rate	Survival_Pct
Q	77	0.38961038961038963	38.96
null	2	1.0	100.0
C	168	0.5535714285714286	55.36
S	644	0.33695652173913043	33.7

EDA Complete - Followed up on all Assignment 1 questions!

```
=====
```

### STEP 3: FEATURE ENGINEERING

```
=====
```

```
[1] Age: Filled with median by class and gender  
[2] Embarked: Filled with mode  
[3] Fare: Filled with median  
[4] Title: Extracted and grouped  
[5] FamilySize: Created from SibSp + Parch + 1  
[6] IsAlone: Binary indicator  
[7] AgeGroup: Categorized into 5 groups  
[8] FareGroup: Categorized into 4 groups  
[9] Deck: Extracted from cabin  
[10] Dropped: PassengerId, Name, Ticket, Cabin
```

Final Features: Pclass, Sex, Survived, Age, SibSp, Parch, Fare, Embarked, Title, FamilySize, IsAlone, AgeGroup, FareGroup, Deck

Sample Data:

Pclass	Sex	Survived	Age	SibSp	Parch	Fare	Embarked	Title	FamilySize	IsAlone	AgeGroup	FareGroup	Deck
3	male	0	22.0	1	0	7.25	S	Mr	2	0	Young_Adult	Low	Unknown
1	female	1	38.0	1	0	71.2833	C	Mrs	2	0	Adult	Very_High	C
3	female	1	26.0	0	0	7.925	S	Miss	1	1	Young_Adult	Medium	Unknown
1	female	1	35.0	1	0	53.1	S	Mrs	2	0	Young_Adult	Very_High	C
3	male	0	35.0	0	0	8.05	S	Mr	1	1	Young_Adult	Medium	Unknown

only showing top 5 rows

```
=====
STEP 4: MODEL TRAINING AND EVALUATION
=====
```

```
[1] Data Split
```

```
Training Set: 746 records
```

```
Validation Set: 145 records
```

```
[2] MODEL EVALUATION RESULTS
```

```
Accuracy: 85.52%
```

```
Precision: 85.43%
```

```
Recall: 85.52%
```

```
F1 Score: 85.43%
```

```
AUC-ROC: 0.8985
```

```
[3] CONFUSION MATRIX
```

```
-----+-----+-----+
| Survived|prediction|count|
-----+-----+-----+
|        0|      0.0|   81|
|        0|      1.0|    9|
|        1|      0.0|   12|
|        1|      1.0|   43|
-----+-----+-----+
```

=====

↓ GENERATING PREDICTIONS ON TEST SET

=====

🖨️ Predicted Survived: 149

🖨️ Predicted Died: 269

Total: 418

=====

↑ SAVING RESULTS

=====

Submission Preview:

PassengerId	Survived
892	0
893	0
894	0
895	0
896	1
897	0
898	1
899	0
900	1
901	0
902	0
903	0

```
|     897|      0|
|     898|      1|
|     899|      0|
|    900|      1|
|    901|      0|
|    902|      0|
|    903|      0|
|    904|      1|
|    905|      0|
|    906|      1|
|    907|      1|
|    908|      0|
|    909|      0|
|    910|      0|
|    911|      1|
+-----+
only showing top 20 rows
```

```
Results saved to: titanic_submission.csv
Model Validation Accuracy: 85.52%
Predictions saved to: titanic_submission.csv
```

```
EDA: 10 analyses (20 pts)
Feature Engineering: 10 features (30 pts)
Model Accuracy: 85.52% (50 pts)
=====
```

```
Process finished with exit code 0
```