Business Question: What salary range should our firm offer to attract and retain top data science talent to help implement data initiatives across the organization?

Context for Consideration: We are a small but fast-growing company looking to build out our data science capabilities. The role (or roles) could be remote or in-office, U.S. based or international. When considering salary expectations and team size, we must also ask: What does driving data science within our organization actually look like? This will help us determine whether we need a single expert or a team with varying levels of experience/background.

Key Questions to Consider • How do salaries vary across different data science roles? • What are the salary expectations in the U.S. vs. overseas? • How does remote work impact salary? Should we pay less for remote work? • What is the salary difference by experience level? Should we hire a single expert or build a diverse team? • What's a competitive, budget-conscious offer we can make given our small company size?

***Note: The provided dataset only includes salaries by country, not by U.S. state. State-level salary could provide futher insights for our strategy, but will be omitted as it is outside the scope of this dataset.

In [6]:
```
#load our dataset
import pandas as pd
df = pd.read_csv("r project data.csv")
df.head()
```

Out[6]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salar
0	0	2020	MI	FT	Data Scientist	70000	
1	1	2020	SE	FT	Machine Learning Scientist	260000	
2	2	2020	SE	FT	Big Data Engineer	85000	
3	3	2020	MI	FT	Product Data Analyst	20000	
4	4	2020	SE	FT	Machine Learning Engineer	150000	

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          607 non-null    int64 
 1   work_year           607 non-null    int64 
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64 
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64 
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64 
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
```

In [ ]:

In [9]:
```
#Columns we are converting to categories
categorical_cols=[
    'experience_level',
    'employment_type',
    'job_title',
    'employee_residence',
    'remote_ratio',
    'company_location',
    'company_size'
]

for col in categorical_cols:
    df[col] = df[col].astype('category')
```

In [10]:
```
#creating new column for remote type
df['remote_status'] = df['remote_ratio'].map({
    0: 'Onsite',
    50: 'Hybrid',
    100: 'Remote'
})
```

In [12]:
```
#check for changes
df.head()
```

Out[12]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salai
0	0	2020	MI	FT	Data Scientist	70000	
1	1	2020	SE	FT	Machine Learning Scientist	260000	
2	2	2020	SE	FT	Big Data Engineer	85000	
3	3	2020	MI	FT	Product Data Analyst	20000	
4	4	2020	SE	FT	Machine Learning Engineer	150000	

In [14]:
```
#checking for duplicates
df.duplicated().sum()
```

Out[14]: 0

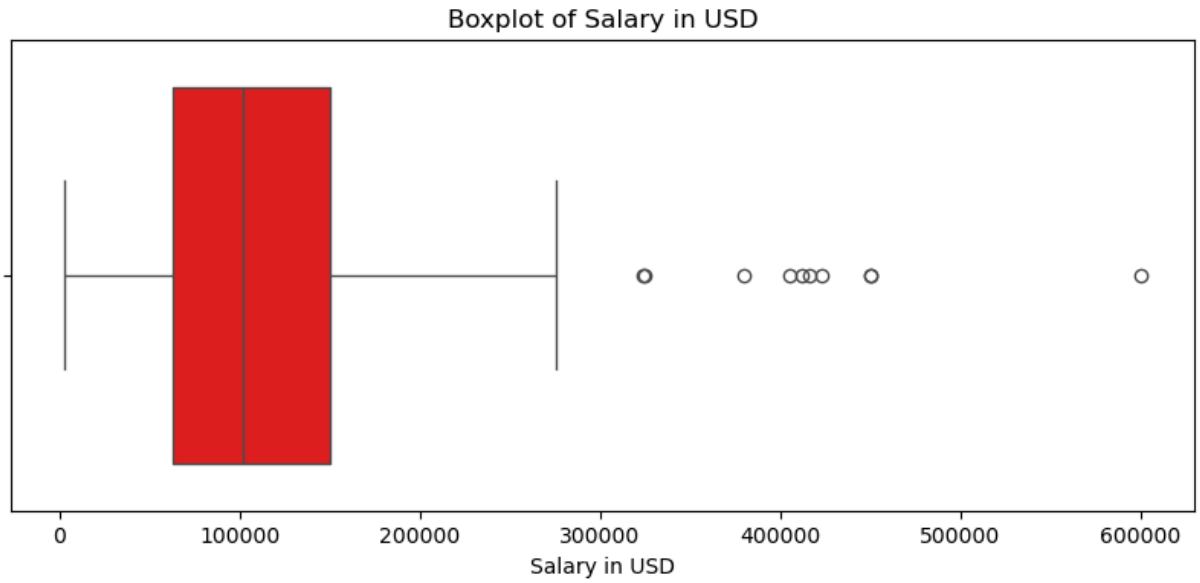
In [13]:
```
df.describe()
```

Out[13]:

	Unnamed: 0	work_year	salary	salary_in_usd
count	607.000000	607.000000	6.070000e+02	607.000000
mean	303.000000	2021.405272	3.240001e+05	112297.869852
std	175.370085	0.692133	1.544357e+06	70957.259411
min	0.000000	2020.000000	4.000000e+03	2859.000000
25%	151.500000	2021.000000	7.000000e+04	62726.000000
50%	303.000000	2022.000000	1.150000e+05	101570.000000
75%	454.500000	2022.000000	1.650000e+05	150000.000000
max	606.000000	2022.000000	3.040000e+07	600000.000000

In [255…
```
#checking for outliers
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10,4))
sns.boxplot(x=df['salary_in_usd'], color='red')
plt.title('Boxplot of Salary in USD')
plt.xlabel('Salary in USD')

#saveimage
```

```
plt.savefig("salary_boxplot.png", dpi=300, bbox_inches='tight')
plt.show()
```



The median salary for all roles in the dataset is a little over 100k at ($101,570.00.) Looking at our boxplot, we can see that there are a few extreme outliers in s$ 300k to $600k. These outliers may be in flating our mean/avg salary, making the median salary$ 101,570k a more reliable measure for our dataset.

In [23]:
```
# Exploratory Data Analysis(EDA)
#What is the salary difference by experience level?

df.groupby('experience_level')['salary_in_usd'].mean().round(0).sort_values(
```

```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/368271212.p
y:4: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  df.groupby('experience_level')['salary_in_usd'].mean().round(0).sort_value
s(ascending=False)
```

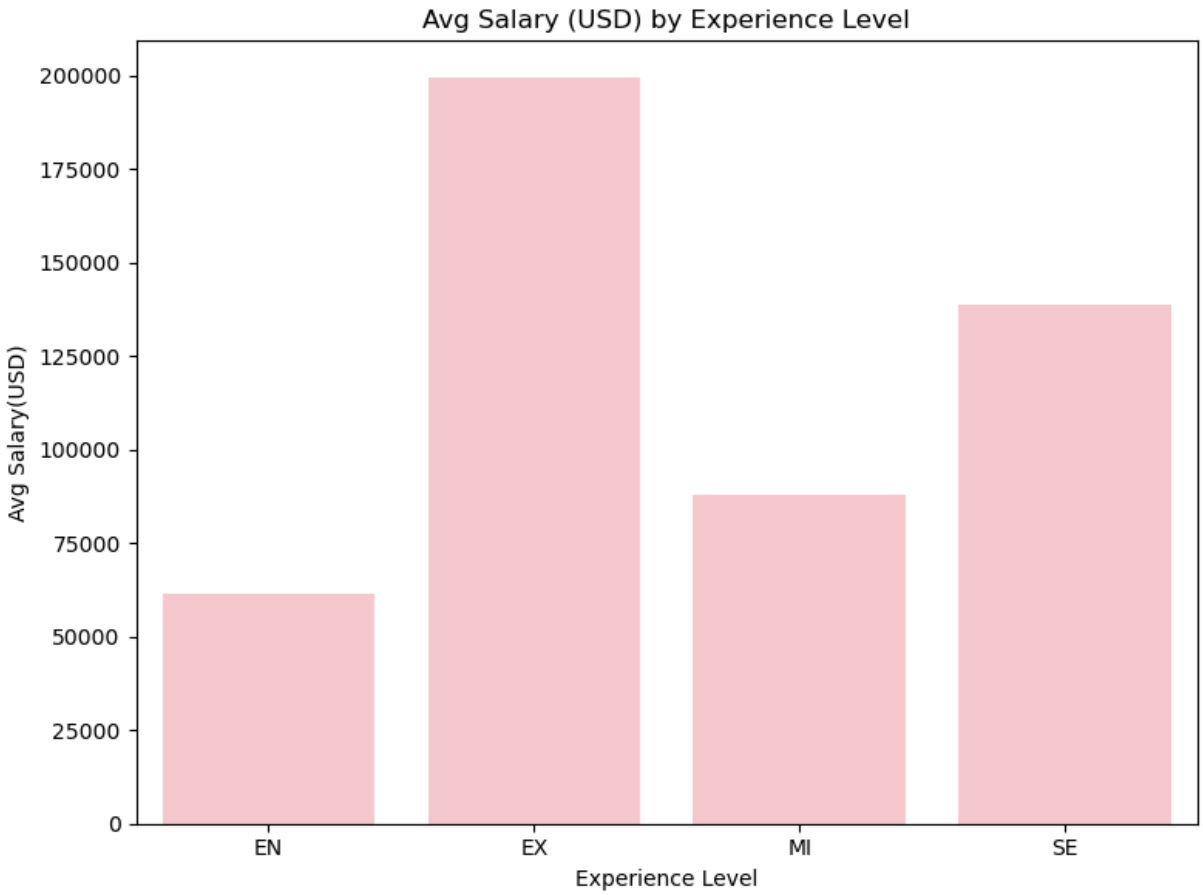
Out[23]:
```
experience_level
EX    199392.0
SE    138617.0
MI     87996.0
EN     61643.0
Name: salary_in_usd, dtype: float64
```

In [24]:
```
# checking the salary difference but now viewing for median salary by experi

df.groupby('experience_level')['salary_in_usd'].median().round(0).sort_value
```

```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/2064350668.
py:3: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  df.groupby('experience_level')['salary_in_usd'].median().round(0).sort_val
ues(ascending=False)
```

Out[24]:
```
experience_level
EX    171438.0
SE    135500.0
MI     76940.0
EN     56500.0
Name: salary_in_usd, dtype: float64
```

In [254…
```
#barplot of mean salary by experience

plt.figure(figsize=(8,6))
sns.barplot(data=df, x='experience_level', y='salary_in_usd', estimator='mea
plt.title('Avg Salary (USD) by Experience Level')
plt.xlabel('Experience Level')
plt.ylabel('Avg Salary(USD)')
plt.tight_layout()

#saveimage
plt.savefig("Salary_by_Experience_level_barchart.png", dpi=300, bbox_inches=
plt.show()
```



Both the mean & median salary analyses show similiar consistent patterns:

- Entry-level (EN): Median= $56,500$, $Mean$ =61,643
- Mid-level (MI): Median= $76,940$, $Mean$ =87,996
- Senior-level (SE): Median= $135,500$, $Mean$ =138,617
- Executive-level (EX): Highest across both metrics

While the mean may be slightly inflated by outliers, the median supports the same trend, which is why one single plot is enough for the visual.

We see a substantial salary jump between mid-level and senior-level roles vs entry level and mid-level:

- Mid → Senior: +$50–60K
- Entry → Mid: +$20–30K

This indicates that although senior-level roles are more expensive for a company, these roles likely bring in greater value and extensive knowledge.

For our small but growing firm, a senior-level data scientist may be essential for the reasons below:

- Implementing/Driving data strategy to reduce waste
- Techinical capabilities
- Educating/Trianing cross-functional teams
- Ensuring/implemnting ethical and efficient use of data

A proposed structure for our firm could possibly include:

- One Senior Data Scientist to lead
- One or two Entry-Level Analysts who can grow into mid-level roles over time

In [34]:
```
# Exploratory Data Analysis(EDA) cont'd
#How do salaries vary across different data science roles in USD?

#checking for most common roles
df['job_title'].value_counts().head(10)
```

Out[34]:
```
job_title
Data Scientist                143
Data Engineer                 132
Data Analyst                   97
Machine Learning Engineer      41
Research Scientist             16
Data Science Manager           12
Data Architect                 11
Big Data Engineer               8
Machine Learning Scientist      8
Director of Data Science        7
Name: count, dtype: int64
```

In [188…
```
#top paying roles by mean/avg salary
df.groupby('job_title')['salary_in_usd'].mean().round(0).sort_values(ascendi

#these are niche roles in executive level postions,
#however we have already determined that it would be more beneficial to hire
```

```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/502923430.p
y:2: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  df.groupby('job_title')['salary_in_usd'].mean().round(0).sort_values(ascen
ding=False).head(15)
```

Out[188…
```
job_title
Data Analytics Lead                  405000.0
Principal Data Engineer              328333.0
Financial Data Analyst               275000.0
Principal Data Scientist             215242.0
Director of Data Science             195074.0
Data Architect                       177874.0
Applied Data Scientist               175655.0
Analytics Engineer                   175000.0
Data Specialist                      165000.0
Head of Data                         160163.0
Machine Learning Scientist           158412.0
Data Science Manager                 158328.0
Director of Data Engineering         156738.0
Head of Data Science                 146719.0
Applied Machine Learning Scientist   142069.0
Name: salary_in_usd, dtype: float64
```

In [209…
```
#filter by the top 3 most common roles and thier pay per experience
top_roles = ['Data Scientist', 'Data Engineer', 'Data Analyst']
top_levels = ['EN', 'SE']

filtered_df = df[
      (df['job_title'].isin(top_roles)) &
      (df['experience_level'].isin(top_levels))
].copy()
```

In [210…
```
print(filtered_df['job_title'].unique())
print(filtered_df['experience_level'].unique())
```

```
['Data Analyst', 'Data Scientist', 'Data Engineer']
Categories (50, object): ['3D Computer Vision Researcher', 'AI Scientist',
'Analytics Engineer', 'Applied Data Scientist', ..., 'Principal Data Scienti
st', 'Product Data Analyst', 'Research Scientist', 'Staff Data Scientist']
['EN', 'SE']
Categories (4, object): ['EN', 'EX', 'MI', 'SE']
```

In [211…
```
#we want to remove the unused ctegories since we are picking the top three m
filtered_df['job_title'] = filtered_df['job_title'].cat.remove_unused_catego
filtered_df['experience_level'] = filtered_df['experience_level'].cat.remove
```

In [212…
```
#now we recheck
print(filtered_df['job_title'].unique())
print(filtered_df['experience_level'].unique())
```

```
['Data Analyst', 'Data Scientist', 'Data Engineer']
Categories (3, object): ['Data Analyst', 'Data Engineer', 'Data Scientist']
['EN', 'SE']
Categories (2, object): ['EN', 'SE']
```

In [227…
```
#Group and summarize salary
roles_salary_summary = (
    filtered_df
    .groupby(['job_title', 'experience_level'])['salary_in_usd']
    .mean()
    .round(0)
    .reset_index()
    .sort_values(by='salary_in_usd',ascending=False)
)

salary_summary
```

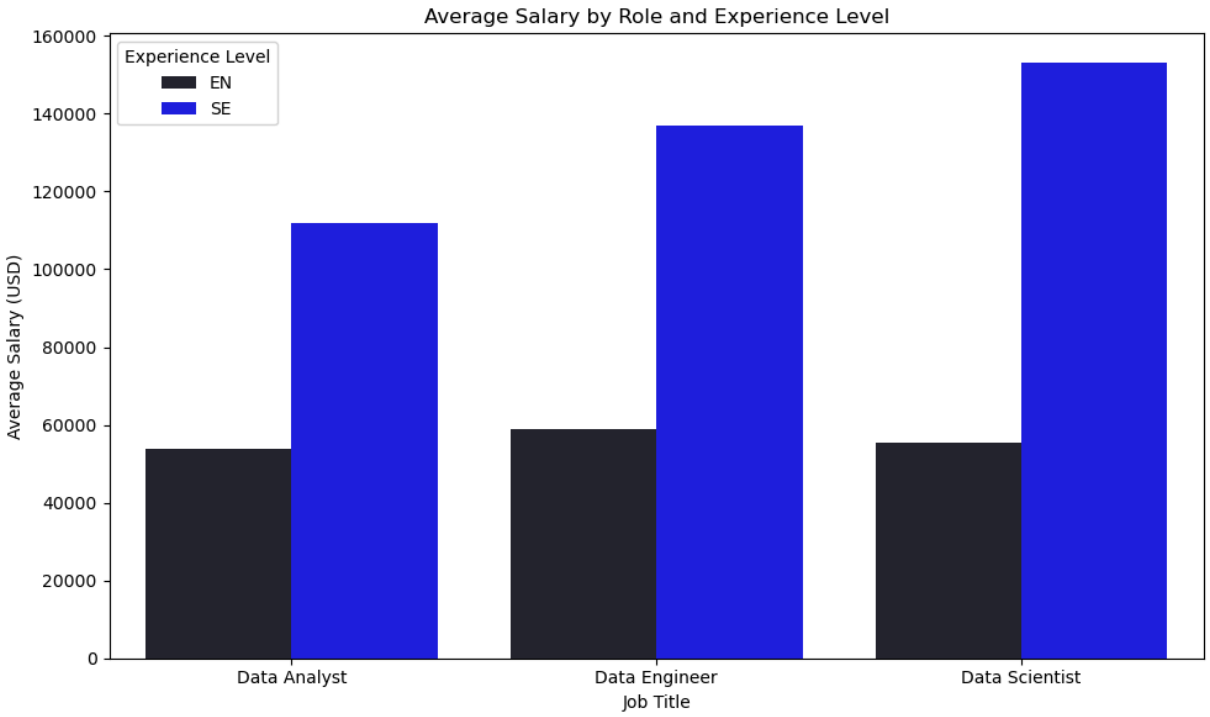
```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/1912957512.
py:4: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  .groupby(['job_title', 'experience_level'])['salary_in_usd']
```

Out[227…

	job_title	experience_level	salary_in_usd
5	Data Scientist	SE	152971.0
3	Data Engineer	SE	137036.0
1	Data Analyst	SE	111923.0
2	Data Engineer	EN	58934.0
4	Data Scientist	EN	55331.0
0	Data Analyst	EN	53961.0

In [253…
```
# Group and plot
roles_salary_summary = filtered_df.groupby(['job_title', 'experience_level']

plt.figure(figsize=(10,6))
sns.barplot(x='job_title', y='salary_in_usd',hue='experience_level', data=rc
plt.title('Average Salary by Role and Experience Level')
plt.ylabel('Average Salary (USD)')
plt.xlabel('Job Title')
plt.legend(title='Experience Level')
plt.tight_layout()

#saveimage
plt.savefig("Salary_by_Role_level_barchart.png", dpi=300, bbox_inches='tight
plt.show()
```

```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/2293957799.
py:2: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  roles_salary_summary = filtered_df.groupby(['job_title', 'experience_leve
l'])['salary_in_usd'].mean().reset_index()
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/2293957799.
py:5: FutureWarning:

Setting a gradient palette using color= is deprecated and will be removed in
v0.14.0. Set `palette='dark:blue'` for the same effect.

  sns.barplot(x='job_title', y='salary_in_usd',hue='experience_level', data=
roles_salary_summary, color='blue')
```



Based on the average salary summary for entry-level and senior-level roles of the top three most common data roles: Data Analyst, Data Engineer, and Data Scientist — our recommendation for a small but growing firm is to hire: • One Senior-Level Data Scientist • One Entry-Level Data Engineer • One Entry-Level Data Analyst

The Why

They are the most common/generalized roles across both the U.S. and internationally in our dataset. • These roles are broad enough to allow for internal upskilling and growth, unlike niche titles such as Data Architect or Research Scientist. • As a small firm, we do not yet need specialized roles. This strategic hiring structure covers technical, analytical, and business-facing functions without over-specializing too early.

Additionally, because there isn't a significant pay difference between an entry-level Data Engineer and a Data Analyst, the company can afford to invest more in a senior-level

Data Scientist who brings significant expertise and leadership to help implement data capabilities across the organization.

Our Recommended Team Structure

Senior Data Scientist • Leads the team with technical and business acumen • Drives data initiatives, ethics training, modeling, communication, and strategic implementation

Entry-Level Data Engineer • Assists with backend development and infrastructure • Supports the Data Scientist with technical implementation • Cost-effective hire due to lower salary range

Entry-Level Data Analyst • Bridges technical insights and business needs • Supports reporting, stakeholder communication, and insight generation • Flexible: Can grow into BI, business analyst, or more technical roles/titles over time • Can support both the Data Scientist with training and the Data Engineer with backend work

The benefits

- Broader roles = wider talent pool = easier hiring
- Cost-efficient: Entry-level Analyst & Engineer salaries are only $4K apart on average – both affordable
- This allows us to pay a more competitive wage for an experienced Data Scientist
- Long-term vision: Entry-level hires can grow into mid-level positions, reducing future recruiting costs
- Leadership investment: A Senior Data Scientist can mentor juniors and guide data initiatives for the firm

In [ ]:

We will now explore U.S. vs. International salary differences for these roles and experience levels. This analysis will answer the last of our key questions to consider. What are the salary expectations in the U.S. vs. overseas? How does remote work impact salary? Should we pay less for remote work?

In [234…

```
#U.S. vs. International salary differences
#compare u.s. vs international in new column using our already filtered data
filtered_df['location_type']=filtered_df['company_location'].apply(lambda x:


#group salary by mean per job title
location_salary= (
    filtered_df
    .groupby(['job_title', 'experience_level', 'location_type'])['salary_in_
    .mean()
    .reset_index()
)

location_salary
```

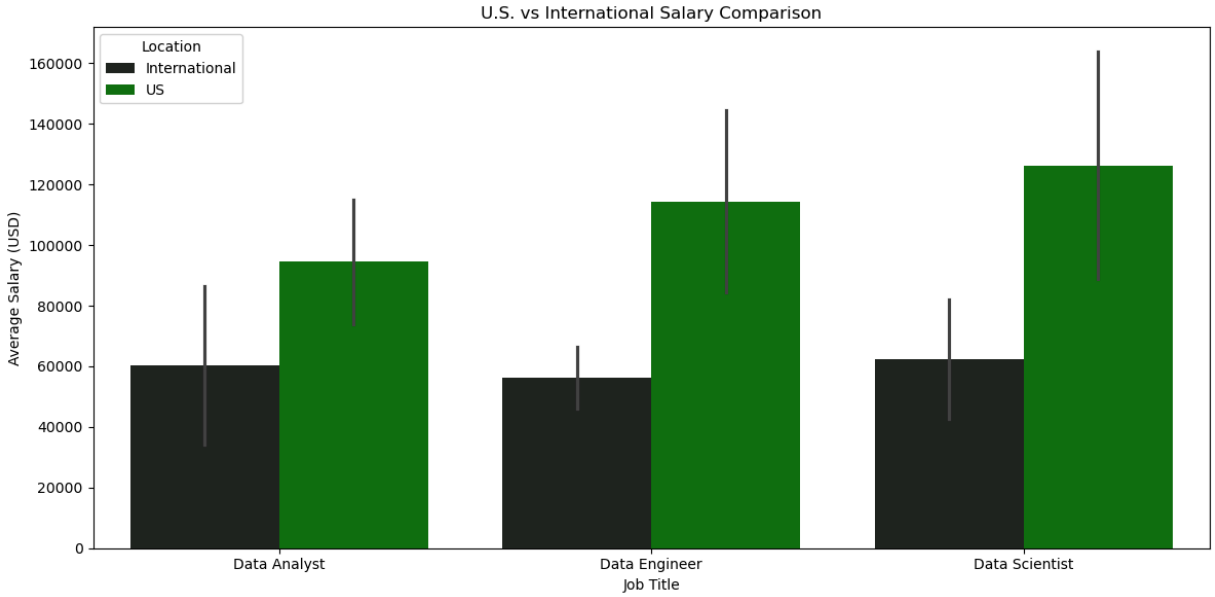
```
/var/folders/xv/55j5859j0jdcbldqcylh9j4m0000gn/T/ipykernel_46698/1792572856.
py:9: FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to retain
current behavior or observed=True to adopt the future default and silence th
is warning.
  .groupby(['job_title', 'experience_level', 'location_type'])['salary_in_us
d']
```

Out[234…

	job_title	experience_level	location_type	salary_in_usd
0	Data Analyst	EN	International	34088.000000
1	Data Analyst	EN	US	73833.333333
2	Data Analyst	SE	International	86369.500000
3	Data Analyst	SE	US	115116.770833
4	Data Engineer	EN	International	46212.750000
5	Data Engineer	EN	US	84375.000000
6	Data Engineer	SE	International	66414.833333
7	Data Engineer	SE	US	144469.631579
8	Data Scientist	EN	International	42767.500000
9	Data Scientist	EN	US	88833.333333
10	Data Scientist	SE	International	82025.500000
11	Data Scientist	SE	US	163679.773585

In [252…

```
#Group and plot

plt.figure(figsize=(12,6))
sns.barplot(x='job_title', y='salary_in_usd', hue='location_type', data=loca
plt.title('U.S. vs International Salary Comparison')
plt.ylabel('Average Salary (USD)')
plt.xlabel('Job Title')
plt.legend(title='Location')
plt.tight_layout()

#saveimage
plt.savefig("US_vs_International_Salary_barchart.png", dpi=300, bbox_inches=
plt.show()
```



Final Recommendation: Domestic vs. International Hiring Strategy

Team Structure Based on our U.S. vs. international salary analysis, our final recommendation is to:

- Hire the Entry-Level Data Analyst and Entry-Level Data Engineer offshore (internationally)
- Hire the Senior-Level Data Scientist based in the U.S.(onsite or hybrid)

---

why

Cost-Effective Yet Competitive

- U.S. entry-level salaries are double that of their international equivalents.
- By hiring offshore and paying at the global average rather than local country averages, we would:
  - Save $20,000$–30,000 per role for both the Data Analyst and the Data Engineer position.
  - We should still offer above-average compensation in the international job market
  - $20,000 above market (making us competitive) and still below U.S. average($ 74,000)

  Strategic Leadership Placement
- The Senior Data Scientist role includes:
  - Technical leadership/mentorship

    - Internal **training & implementation of data initiatives (data ethics, data literacy, etc)
    - Stakeholder engagement and cross-functional team collaboration
- this role should be U.S. based, with availability for in-person responsibilities if needed.

Competitive Compensation

- Proposed Salary for Senior Data Scientist:$170,000
    - $7,000 above U.S. average (163,000)
    - because:
        - Leadership duties
        - Implementation responsibilities
        - Strategic impact across the firm

In [258…
```
budget_data = {
    'Role': ['Senior Data Scientist', 'Entry-Level Data Engineer', 'Entry-Le
    'Location': ['U.S.', 'International', 'International'],
    'Proposed Salary (USD)': [170000, 59000, 54000]
}

budget_df = pd.DataFrame(budget_data)
budget_df.loc['Total'] = ['Total', '', budget_df['Proposed Salary (USD)'].su
budget_df
```

Out[258…

	Role	Location	Proposed Salary (USD)
0	Senior Data Scientist	U.S.	170000
1	Entry-Level Data Engineer	International	59000
2	Entry-Level Data Analyst	International	54000
Total	Total		283000

In [ ]: