1. Importing Dependencies

In [1]:
```
import pandas as pd
import numpy as np
import altair as alt
```

In [2]:
```
alt.data_transformers.disable_max_rows()
```

Out[2]: DataTransformerRegistry.enable('default')

2. Importing Processed Data

In [3]:
```
df = pd.read_csv('processed_data.csv')
df.columns = [c.strip().lower().replace(' ', '_') for c in df.columns]
df.shape
```

Out[3]: (30000, 29)

3. Feature Engineering

In [4]:
```
# Standardizing column names for consistency
df.columns = [col.strip().lower() for col in df.columns]
```

In [5]:
```
# Credit Utilization: how much of the limit is being used
df['utilization'] = (df['bill_amt1'] / df['limit_bal'].replace(0, np.nan)
```

In [6]:
```
# Payment Summaries and Billing of 6 Months
bill_cols = [f'bill_amt{i}' for i in range(1, 7)]
pay_cols  = [f'pay_amt{i}' for i in range(1, 7)]
df['avg_bill_6m']  = df[bill_cols].mean(axis=1)
df['total_pay_6m'] = df[pay_cols].sum(axis=1)
```

In [7]:
```
# Billing Trend: upward or downward over 6 months
months = np.arange(1, 7)
def slope_row(r):
    y = r[bill_cols].values.astype(float)
    valid = ~np.isnan(y)
    if valid.sum() < 3:
        return np.nan
    return np.polyfit(months[valid], y[valid], 1)[0]
df['slope'] = df[bill_cols].apply(slope_row, axis=1)
```

In [8]:
```
# Normalizing main indicators
def minmax_s(x):
    s = x.fillna(0).astype(float)
    return (s - s.min()) / (s.max() - s.min())

df['util_norm']  = minmax_s(df['utilization'])
df['slope_norm'] = minmax_s(df['slope'].clip(lower=0))
df['bill_norm']  = minmax_s(df['avg_bill_6m'].clip(lower=0))
```

In [9]:
```
# Composite Risk Score (weighted mix)
w_util, w_slope, w_bill = 0.5, 0.3, 0.2
df['risk_score'] = (
    w_util  * df['util_norm'] +
    w_slope * df['slope_norm'] +
    w_bill  * df['bill_norm']
).clip(0, 1)
```

In [10]:
```
# Risk Categories
df['risk_flag'] = pd.cut(
    df['risk_score'],
    bins=[-0.01, df['risk_score'].quantile(0.7),
          df['risk_score'].quantile(0.9), 1.01],
    labels=['Low', 'Medium', 'High'],
    include_lowest=True
)
```

In [11]:
```
df[['utilization', 'avg_bill_6m', 'total_pay_6m', 'slope', 'risk_score',
```

Out[11]:

	utilization	avg_bill_6m	total_pay_6m	slope	risk_score	risk_flag
0	0.195650	1284.000000	689.0	-844.571429	0.015447	Low
1	0.022350	2846.166667	5000.0	247.857143	0.003128	Low
2	0.324878	16942.166667	11018.0	-1854.714286	0.029026	Low
3	0.939800	38555.666667	8388.0	-4743.257143	0.081582	Medium
4	0.172340	18223.166667	59049.0	2231.514286	0.024240	Low

4. Risk Analytics

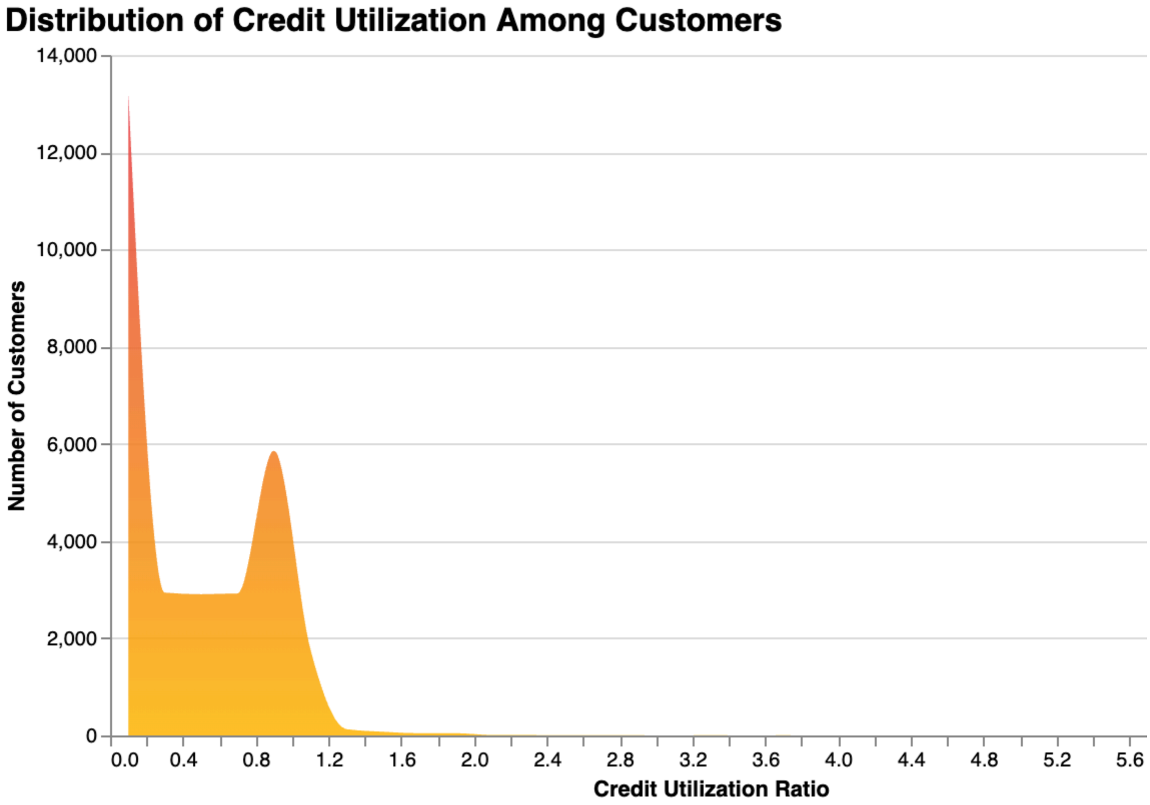
A. Distribution of Credit Utilization Among Customers

In [14]:
```
util_plot = alt.Chart(df).mark_area(
    interpolate='monotone',
    color=alt.Gradient(
        gradient='linear',
        stops=[
            alt.GradientStop(color='#FFB703', offset=0),
            alt.GradientStop(color='#E63946', offset=1)
        ],
        x1=1, x2=1, y1=1, y2=0
```

```
    ),
    opacity=0.85
).encode(
    alt.X('utilization:Q', bin=alt.Bin(maxbins=40), title='Credit Utiliza
    alt.Y('count()', title='Number of Customers')
).properties(
    title='Distribution of Credit Utilization Among Customers',
    width=600,
    height=340
).configure_title(fontSize=16, anchor='start')

util_plot
```

Out[14]:



This chart shows how customers are using their available credit limits.
Each bar represents a group of customers based on their **credit utilization ratio**, which is the proportion of their credit limit that they've used.

- Most customers stay within moderate utilization levels, which indicates healthy financial behavior.
- A smaller number of customers show very high utilization which is a possible sign of credit stress or dependency.
- The overall shape of the distribution helps us understand the general spending and repayment balance in the customer base.

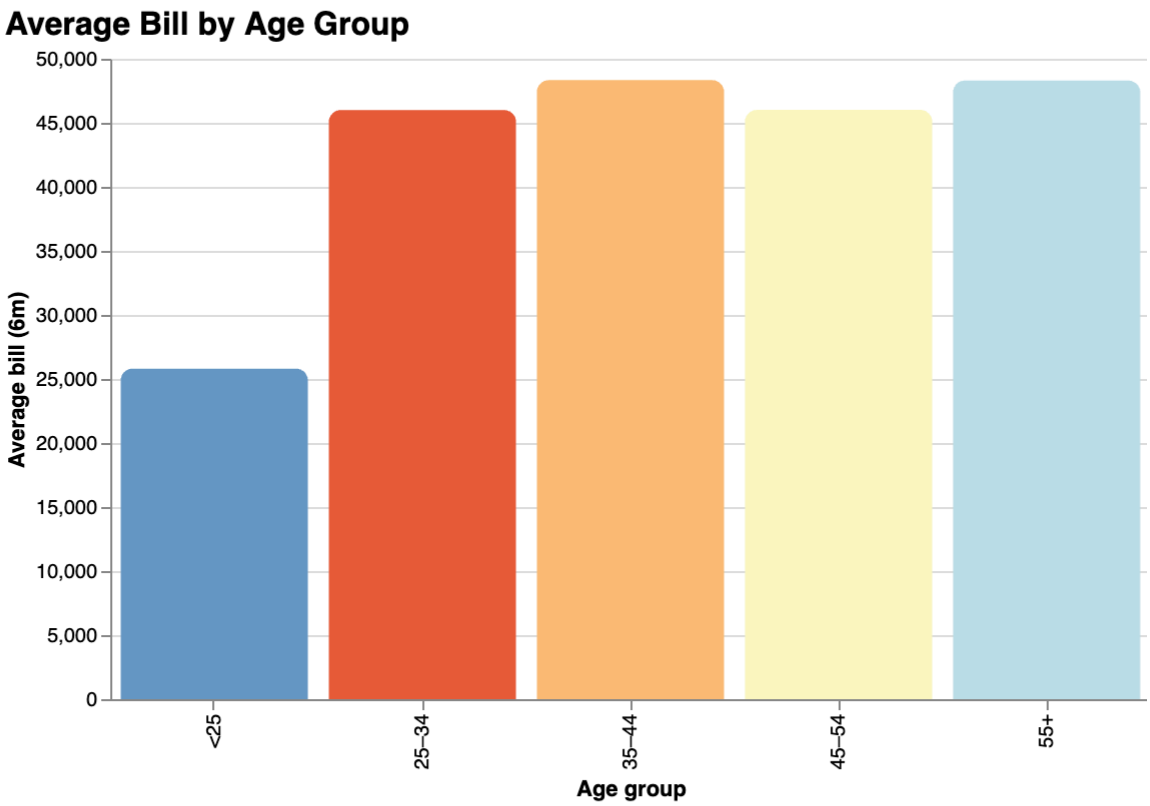
In short, **lower utilization** often reflects better financial management, while **high utilization** may require closer monitoring.

B. Average Bill by Age Group

In [15]:
```
# Creating age group labels
if 'age_group' not in df.columns:
    df['age_group'] = pd.cut(
        df['age'],
        bins=[0, 24, 34, 44, 54, df['age'].max()],
        labels=['<25', '25–34', '35–44', '45–54', '55+']
    )
```

In [16]:
```
age_bill = alt.Chart(df).transform_filter("!isNaN(datum.avg_bill_6m)").ma
    cornerRadiusTopLeft=6,
    cornerRadiusTopRight=6
).encode(
    x=alt.X('age_group:N', sort=['<25','25–34','35–44','45–54','55+'], ti
    y=alt.Y('mean(avg_bill_6m):Q', title='Average bill (6m)'),
    color=alt.Color('age_group:N', scale=alt.Scale(scheme='redyellowblue'
    tooltip=[
        alt.Tooltip('mean(avg_bill_6m):Q', title='Avg bill (6m)', format=
        'age_group'
    ]
).properties(
    title='Average Bill by Age Group',
    width=520,
    height=320
).configure_title(fontSize=16, anchor='start')
age_bill
```

Out[16]:



This bar chart compares **average monthly bill amounts** across different age groups.

- Younger customers tend to have smaller average bills, which could reflect lower incomes or lighter credit usage.

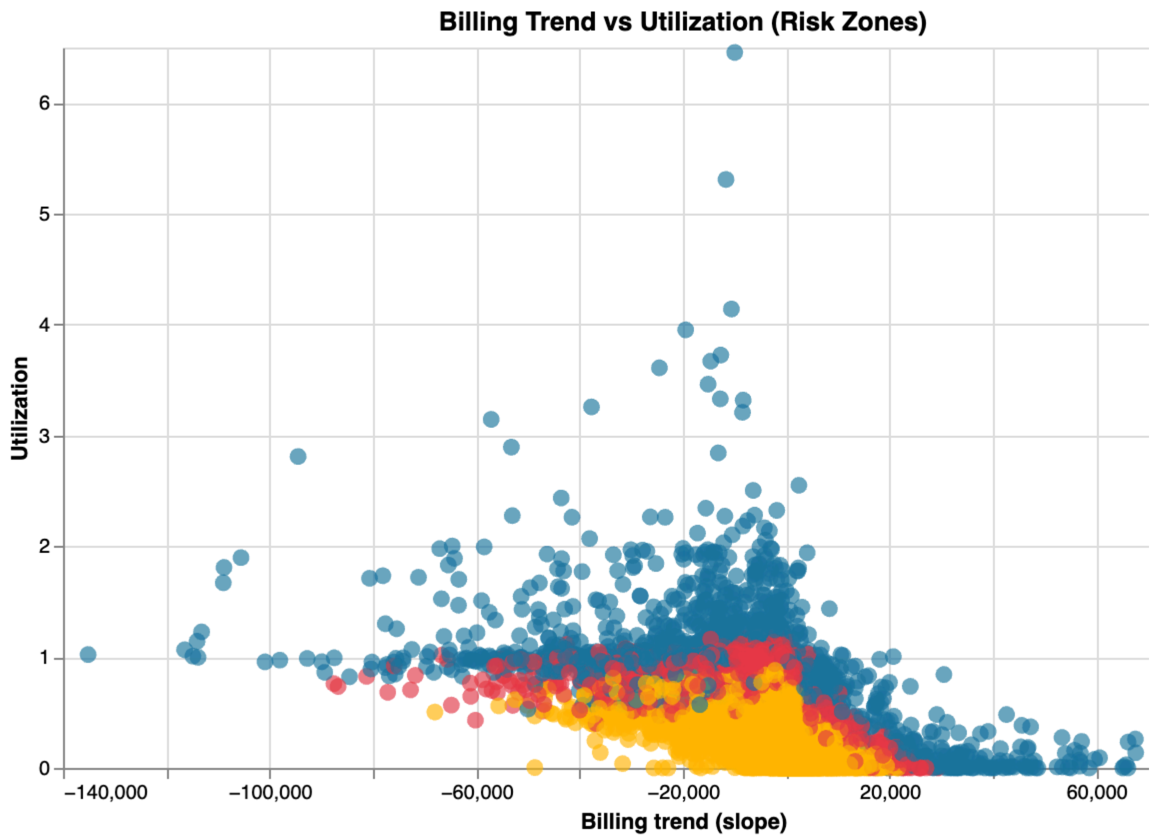
- Middle-aged and older customers often show higher bill amounts, possibly due to higher credit access or spending needs.
- The differences across age groups help highlight spending habits and potential customer segmentation opportunities.

Overall, this view offers insight into how **age influences billing behavior** and financial engagement.

C. Billing Trend vs Utilization

In [17]:
```
slope_util = alt.Chart(df).transform_filter("!isNaN(datum.slope) && !isNa
    x=alt.X('slope:Q', title='Billing trend (slope)'),
    y=alt.Y('utilization:Q', title='Utilization'),
    color=alt.Color('risk_flag:N', scale=alt.Scale(domain=['Low','Medium'
    tooltip=['slope','utilization','risk_flag']
).properties(title='Billing Trend vs Utilization (Risk Zones)', width=620
slope_util
```

Out[17]:



This scatter plot examines the relationship between customers' **billing trends** and their **credit utilization** levels.

- Customers with **rising bills and high utilization** are grouped toward the higher-risk zones.
- Customers with **stable or decreasing bills** and moderate utilization usually fall in lower-risk zones.

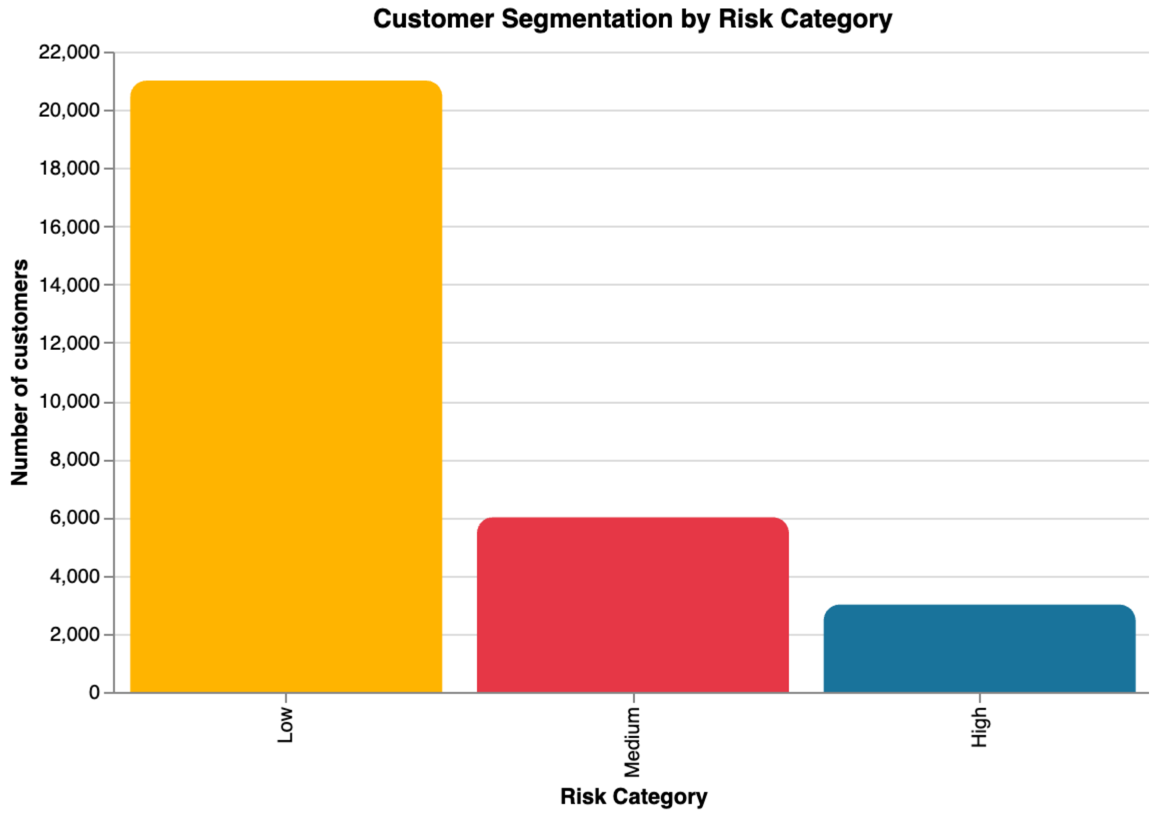
- The chart uses color-coded zones to make it easier to identify risk behavior visually.

This visualization helps connect **spending growth** and **credit reliance** to overall financial risk.

D. Customer Segmentation by Risk Category

In [18]:
```
risk_segment = alt.Chart(df).mark_bar(cornerRadiusTopLeft=8, cornerRadius
    x=alt.X('risk_flag:N', title='Risk Category', sort=['Low','Medium','H
    y=alt.Y('count():Q', title='Number of customers'),
    color=alt.Color('risk_flag:N', scale=alt.Scale(domain=['Low','Medium'
    tooltip=['risk_flag', 'count()']
).properties(title='Customer Segmentation by Risk Category', width=520, h
risk_segment
```

Out[18]:



This bar chart breaks down the customer base into **Low**, **Medium**, and **High** risk segments.

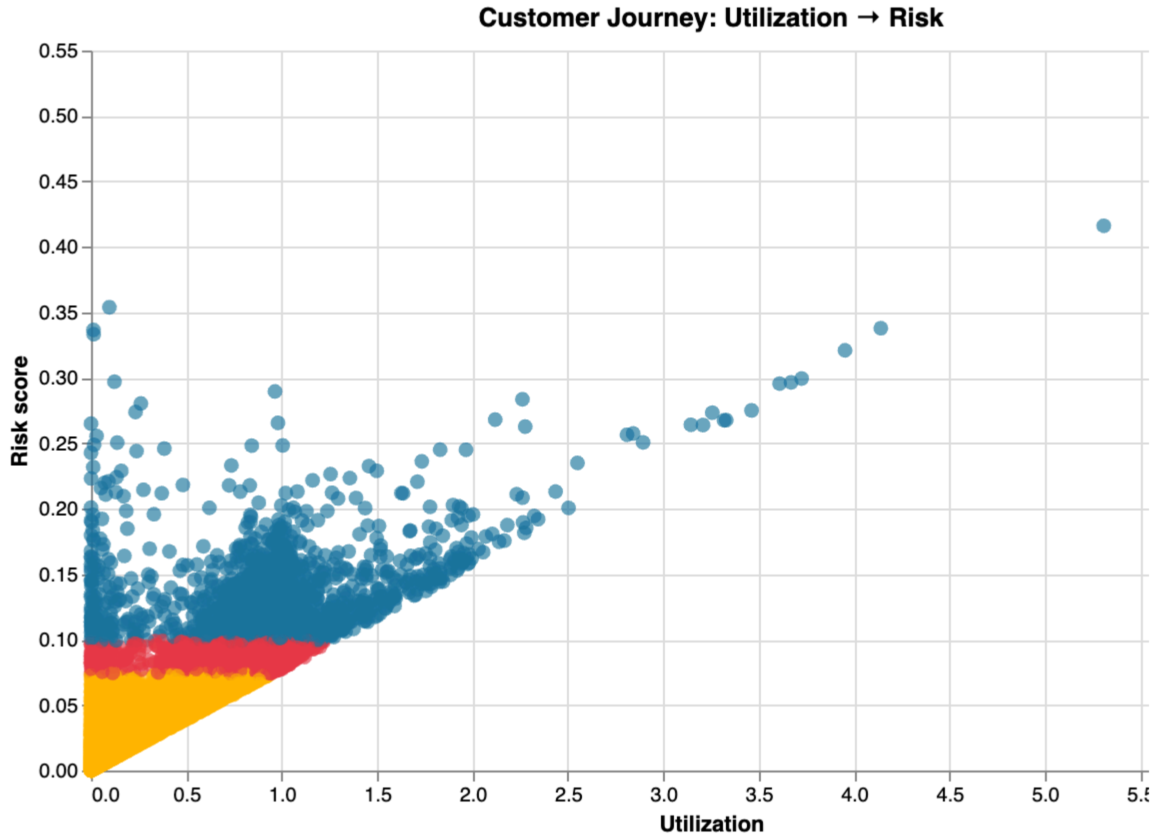
- A larger share of low-risk customers indicates strong financial health across the dataset.
- The presence of medium- and high-risk groups signals areas that may need early intervention or closer analysis.
- Each bar reflects the number of customers falling within that specific risk band.

Such segmentation helps institutions focus attention and resources where they're needed most.

E. Customer Journey: Utilization → Risk

In [19]:
```
journey = alt.Chart(df).transform_filter("!isNaN(datum.utilization) && !i
    x=alt.X('utilization:Q', title='Utilization'),
    y=alt.Y('risk_score:Q', title='Risk score'),
    color=alt.Color('risk_flag:N', scale=alt.Scale(domain=['Low','Medium'
    tooltip=['utilization','risk_score','risk_flag']
).properties(title='Customer Journey: Utilization → Risk', width=620, hei
journey
```

Out[19]:



This scatter chart shows how a customer's **credit utilization** connects to their **risk score**.

- As utilization increases, risk scores also tend to rise.
- Low utilization usually aligns with low risk, while very high utilization correlates with high risk.
- The color of each point represents the customer's risk category — providing a quick view of how spending behavior translates into risk levels.

In essence, this visualization illustrates how **credit usage patterns lead to changes in overall financial risk**.

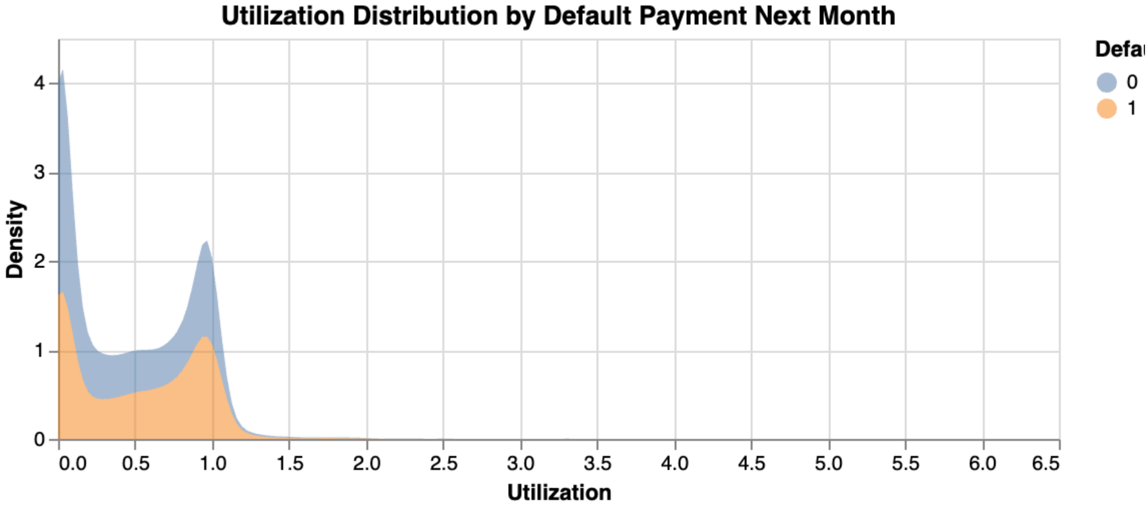
F. Utilization Distribution by Default Payment Next Month

In [20]:
```
target = 'default_payment_next_month' if df['default_payment_next_month']

dens = alt.Chart(df).transform_density(
    'utilization',
    as_=['utilization','density'],
    groupby=[target]
).mark_area(opacity=0.5).encode(
    x=alt.X('utilization:Q', title='Utilization'),
    y=alt.Y('density:Q', title='Density'),
    color=alt.Color(f'{target}:N', title=target.replace('_',' ').title())
).properties(
    width=500, height=200, title=f'Utilization Distribution by {target.re
)

dens
```

Out[20]:



This chart compares the credit utilization levels of customers who **defaulted on their next payment** versus those who did not.

G. Avg Bill per Month by Risk Flag

In [21]:
```
bill_long = df[['id','risk_flag'] + bill_cols].melt(
    id_vars=['id','risk_flag'],
    value_vars=bill_cols,
    var_name='bill_col',
    value_name='bill_amt'
)
bill_long['month_idx'] = bill_long['bill_col'].str.extract('(\d+)').astyp

avg_ts = (bill_long
          .groupby(['risk_flag','month_idx'])
```

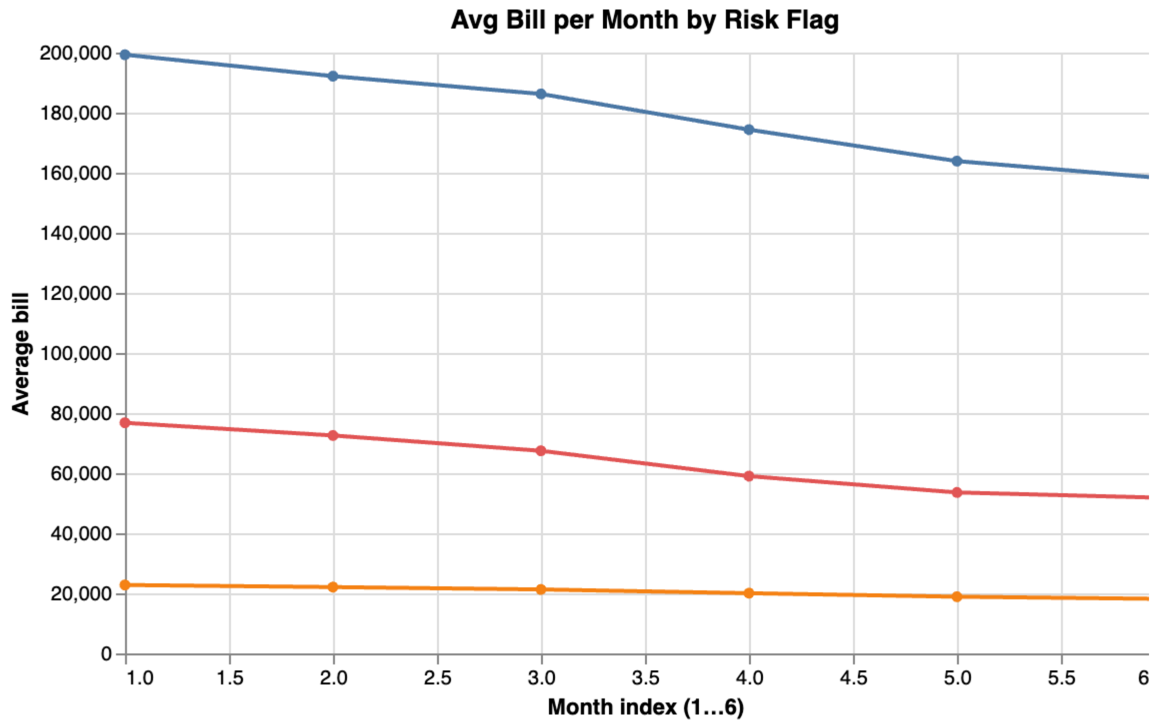
```
            .agg(avg_bill=('bill_amt','mean'))
            .reset_index())

ts = alt.Chart(avg_ts).mark_line(point=True).encode(
    x=alt.X('month_idx:Q', title='Month index (1…6)', scale=alt.Scale(dom
    y=alt.Y('avg_bill:Q', title='Average bill'),
    color=alt.Color('risk_flag:N', title='Risk Flag')
).properties(width=520, height=300, title='Avg Bill per Month by Risk Fla

ts
```

```
<>:7: SyntaxWarning: invalid escape sequence '\d'
<>:7: SyntaxWarning: invalid escape sequence '\d'
/var/folders/nt/4rp641zs5lg7_234f6d1h4_00000gn/T/ipykernel_31289/376789521
5.py:7: SyntaxWarning: invalid escape sequence '\d'
  bill_long['month_idx'] = bill_long['bill_col'].str.extract('(\d+)').asty
pe(int)
/var/folders/nt/4rp641zs5lg7_234f6d1h4_00000gn/T/ipykernel_31289/376789521
5.py:10: FutureWarning: The default of observed=False is deprecated and wi
ll be changed to True in a future version of pandas. Pass observed=False t
o retain current behavior or observed=True to adopt the future default and
silence this warning.
  .groupby(['risk_flag','month_idx'])
```

Out[21]:



This line chart tracks the **average monthly bill amount** over six months for customers in different risk categories.

- Low-risk customers show steady billing patterns across months.
- Medium- and high-risk customers often exhibit higher fluctuations or sharper increases.
- The comparison highlights how consistency in billing behavior tends to align with lower financial risk.

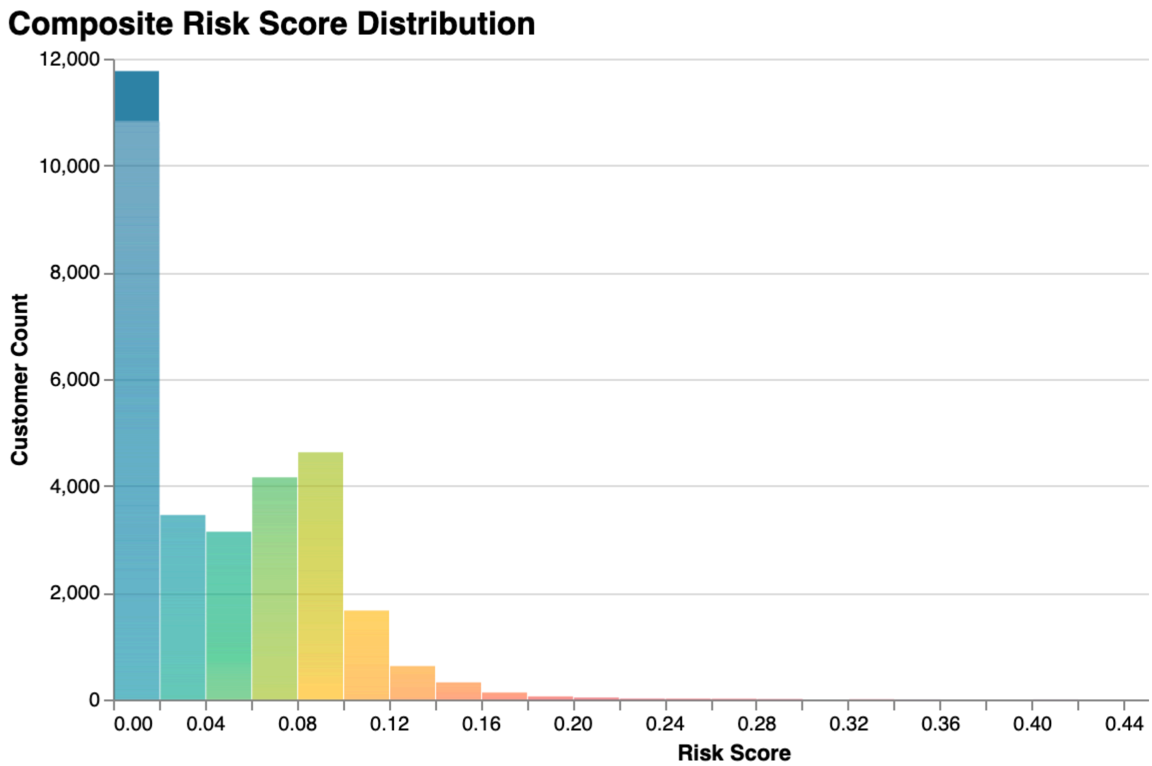
The plot offers a timeline view of **how customer risk correlates with payment stability**.

H. Composite Risk Score Distribution

In [22]:
```
risk_hist = (
    alt.Chart(df)
    .transform_filter(alt.datum.risk_score != None)
    .mark_bar(opacity=0.9)
    .encode(
        x=alt.X('risk_score:Q', bin=alt.Bin(maxbins=35), title='Risk Scor
        y=alt.Y('count():Q', title='Customer Count'),
        color=alt.Color(
            'risk_score:Q',
            scale=alt.Scale(
                domain=[0, 0.1, 0.2],
                range=['#1A759F', '#FFB703', '#E63946'],
                clamp=True
            ),
            title='Risk Score'
        ),
        tooltip=[alt.Tooltip('risk_score:Q', format='.2f'), 'count()']
    )
    .properties(
        title='Composite Risk Score Distribution',
        width=620,
        height=320
    )
    .configure_title(fontSize=16, anchor='start')
)

risk_hist
```

Out[22]:



This histogram shows how customers are distributed across the full range of **risk scores**.

- The gradient smoothly transitions from **low to high risk**, with blue for safer customers and red for riskier ones.
- Most customers cluster toward the lower-risk end, while a small group appears on the higher-risk tail.
- This overall picture helps identify where the majority of the portfolio lies and where attention may be needed.

The visualization provides a final, comprehensive look at the **risk landscape of all customers**.