Pandas & Data Preprocessing - Revision Notes

PANDAS OVERVIEW

Task	Command / Function
Create Series	pd.Series(data, index=labels)
Create DataFrame	pd.DataFrame(dict)
Read CSV	pd.read_csv(‘file.csv')
Inspect Data	df.head(), df.shape, df.info(), df.describe()
Select Column	df['col'], df[['col1','col2']]
Select Row	df.iloc[i], df.loc[label]
Filter Rows	df[(df['col']>x) & (df['col2']=='y')]
Sort Data	df.sort_values(by='col', ascending=False)
Add/Drop Column	df['new']=..., df.drop(‘col', axis=1)
Group & Aggregate	df.groupby(‘col')['val'].mean()
Save File	df.to_csv(‘out.csv', index=False)

DATA PREPROCESSING SUMMARY

Topic	Key Code / Concept
Missing Values	df.isnull().sum(), df.dropna(), df['col'].fillna(df['col'].mean())
Duplicates	df.duplicated().sum(), df.drop_duplicates()
Data Types	df['col']=df['col'].astype(float), pd.to_datetime(df['date'])
Encoding	df['Dept_Code']=df['Dept'].astype(‘category').cat.codes, pd.get_dummies(df, columns=[
Scaling	from sklearn.preprocessing import MinMaxScaler; df[['Age','Salary']] = MinMaxScaler().
Outliers	df['Salary'].describe(), df[df['Salary'] < 200000]
Save Cleaned Data	df.to_csv(‘cleaned_employees.csv', index=False)